



UNIVERSIDAD DE GRANADA
MÁSTER DE CIENCIA DE DATOS E INGENIERÍA DE
COMPUTADORES
CURSO ACADÉMICO 2019-2020
INTRODUCCIÓN A LA CIENCIA DE DATOS

Trabajo integrador

*Análisis y elaboración de modelos descriptivos y/o predictivos
de conjuntos de datos de clasificación y regresión.*

Nicolás Cubero

22 de Diciembre de 2019

Índice

Índice de figuras	5
Índice de tablas	10
1. Introducción	11
2. El dataset <i>house</i> : Problema de regresión.	11
2.1. Descripción del dataset <i>house</i>	11
2.2. Análisis exploratorio del dataset <i>house</i>	13
2.2.1. Análisis de las variables del dataset	13
2.2.2. Estudio de las relaciones entre variables	68
2.3. Elaboración de modelos predictivos	73
2.3.1. Elaboración de modelos lineales simples	73
2.3.2. Elaboración de modelos lineales compuestos	78
2.3.3. Elaboración de modelos basados en k-NN	97
2.3.4. Comparación de modelos	101
3. El dataset <i>vehicle</i> : Problema de clasificación.	106
3.1. Descripción del dataset <i>vehicle</i>	106
3.2. Análisis exploratorio del dataset <i>vehicle</i>	107
3.2.1. Análisis de las variables del dataset	108
3.2.2. Análisis de las relaciones entre las variables	148
3.3. Elaboración de modelos predictivos	150
3.3.1. Elaboración de modelos basados en k-NN	151
3.3.2. Elaboración de modelos basados en LDA	153
3.3.3. Elaboración de modelos basados en QDA	156
3.3.4. Comparación de modelos	158
A. Script1.R	159
B. Script2.R	168
C. Script3.R	185
D. Script4.R	187
E. Script5.R	199

Índice de figuras

1.	Diagrama <i>boxplot</i> de la variable <i>P1</i> de <i>house</i>	16
2.	Diagrama <i>boxplot</i> de la variable <i>P1</i> de <i>house</i> en escala logarítmica en base 10	17
3.	Histograma del subconjunto de la variable <i>P1</i> comprendida en el intervalo $[Q1 - 1,5 * IQR, Q3 + 1,5 * IQR]$	20
4.	Histograma del subconjunto de la variable <i>P1</i> comprendida en el intervalo $(Q3 + 1,5 * IQR, +\infty)$	21
5.	Diagrama <i>boxplot</i> de la variable <i>P5p1</i>	23
6.	Histograma de la variable <i>P5p1</i>	24
7.	Diagrama <i>boxplot</i> de la variable <i>P6p2</i>	26
8.	Histograma de la variable <i>P6p2</i>	27
9.	Histograma de la variable <i>P11p4</i>	29
10.	Histograma de la variable <i>P11p4</i>	30
11.	Histograma de la variable <i>P14p9</i>	32
12.	Histograma de la variable <i>P14p9</i>	33
13.	Histograma de la variable <i>P15p1</i>	35
14.	Histograma de la variable <i>P15p1</i>	36
15.	Histograma de la variable <i>P15p3</i>	38
16.	Histograma de la variable <i>P15p3</i>	39
17.	Histograma de la variable <i>P16p2</i>	41
18.	Histograma de la variable <i>P6p2</i>	42
19.	Histograma de la variable <i>P18p2</i>	44
20.	Histograma de la variable <i>P18p2</i>	45
21.	Histograma de la variable <i>P18p2</i>	46
22.	Histograma de la variable <i>P27p4</i>	48
23.	Histograma de la variable <i>P27p4</i>	49
24.	Histograma de la variable <i>H2p2</i>	51
25.	Histograma de la variable <i>H2p2</i>	52
26.	Histograma de la variable <i>H8p2</i>	54
27.	Histograma de la variable <i>H8p2</i>	55
28.	Histograma de la variable <i>H10p1</i>	57
29.	Histograma de la variable <i>H10p1</i>	58
30.	Histograma de la variable <i>H13p1</i>	60
31.	Histograma de la variable <i>H13p1</i>	61
32.	Representación del diagrama <i>boxplot</i> (diagrama superior) e histograma (diagrama inferior) de la variable <i>H18pA</i>	63
33.	Representación del diagrama <i>boxplot</i> (diagrama superior) e histograma (diagrama inferior) de la variable <i>H40p4</i>	65

34. Representación del diagrama <i>boxplot</i> (diagrama superior) e histograma (diagrama inferior) de la variable <i>Price</i>	67
35. A la derecha se muestra una matriz triangular inferior donde cada celda muestra el valor del coeficiente de Kendall asociado a cada par de variables, nótese que un color cercano a azul oscuro indica una correlación lineal directa, mientras que un color rojo indica una correlación lineal inversa. A la izquierda se representa otro correlograma donde el valor del coeficiente ha sido sustituido por una esfera cuyo color y forma muestran el grado de correlación lineal entre las variables.	71
36. Diagrama de puntos que relaciona la variable P5p1 frente a P1	72
37. Representaciones de la variable <i>Price</i> frente a cada una de las variables independientes del <i>dataset</i> . Para cada una de las representaciones, se representa la nube de puntos en color rojo y la línea que representa el modelo lineal generado en cada caso en color azul claro	76
38. Histograma de la variable <i>Compactness</i>	110
39. Histograma de la variable <i>Compactness</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	111
40. Diagrama <i>boxplot</i> de la variable <i>Circularity</i>	113
41. Histograma de la variable <i>Circularity</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	114
42. Diagrama <i>boxplot</i> de la variable <i>Distance_circularity</i>	116
43. Histograma de la variable <i>Distance_circularity</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	117
44. Diagrama <i>boxplot</i> de la variable <i>Radius_ratio</i>	119
45. Histograma de la variable <i>Radius_ratio</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	120
46. Diagrama <i>boxplot</i> de la variable <i>Praxis_aspect_ratio</i>	123
47. Histograma de la variable <i>Praxis_aspect_ratio</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	123
48. Diagrama <i>boxplot</i> de la variable <i>Max_length_aspect_ratio</i>	125
49. Histograma de la variable <i>Max_length_aspect_ratio</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	126
50. Diagrama <i>boxplot</i> de la variable <i>Scatter_ratio</i>	127

51. Histograma de la variable <i>Scatter_ratio</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	128
52. Diagrama <i>boxplot</i> de la variable <i>Elongatedness</i>	129
53. Histograma de la variable <i>Elongatedness</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	130
54. Diagrama <i>boxplot</i> de la variable <i>Praxis_rectangular</i>	131
55. Histograma de la variable <i>Praxis_rectangular</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	131
56. Diagrama <i>boxplot</i> de la variable <i>Length_rectangular</i>	133
57. Histograma de la variable <i>Length_rectangular</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	133
58. Diagrama boxplot de la variable <i>Major_variance</i>	134
59. Histograma de la variable <i>Major_variance</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	135
60. Diagrama <i>boxplot</i> de la variable <i>Minor_variance</i>	136
61. Histograma de la variable <i>Minor_variance</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	137
62. Diagrama boxplot de la variable <i>Gyration_radius</i>	138
63. Histograma de la variable <i>Gyration_radius</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	139
64. Diagrama boxplot de la variable <i>Major_skewness</i>	140
65. Histograma de la variable <i>Major_skewness</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	140
66. Diagrama <i>boxplot</i> de la variable <i>Minor_skewness</i>	142
67. Histograma de la variable <i>Minor_skewness</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	142
68. Diagrama boxplot de la variable <i>Minor_kurtosis</i>	143
69. Histograma de la variable <i>Minor_kurtosis</i> , se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	144
70. Diagrama boxplot de la variable <i>Major_kurtosis</i>	145

71.	Histograma de la variable Major_kurtosis, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	145
72.	Diagrama boxplot de la variable Hollows_ratio	146
73.	Histograma de la variable Hollows_ratio, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase	147
74.	Diagrama de barras de la variable <i>Class</i>	148
75.	A la derecha se muestra una matriz triangular inferior donde cada celda muestra el valor del coeficiente de <i>Kendall</i> asociado a cada par de variables, nótese que un color cercano a azul oscuro indica una correlación lineal directa, mientras que un color rojo indica una correlación lineal inversa. A la izquierda se representa otro correlograma donde el valor del coeficiente ha sido sustituido por una esfera cuyo color y forma muestran el grado de correlación lineal entre las variables.	150
76.	Cuadro con correlograma que analiza la correlación lineal entre las variales haciendo una separación por clase, como medida de correlación se ha usado el coeficiente de <i>Kendall</i>	157

Índice de cuadros

1.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo P1	15
2.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> caclulados para el atributo <i>P5p1</i>	22
3.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> caclulados para el atributo <i>P6p2</i>	25
4.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> caclulados para el atributo <i>P11p4</i>	28
5.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> caclulados para el atributo <i>P14p9</i>	31
6.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> caclulados para el atributo <i>P15p1</i>	34
7.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> caclulados para el atributo <i>P15p3</i>	37
8.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>P16p2</i>	40
9.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>P18p2</i>	43
10.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>P27p4</i>	47
11.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>H2p2</i>	50
12.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>H8p2</i>	53
13.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>H10p1</i>	56
14.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>H13p1</i>	59
15.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>H18pA</i>	62
16.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>H40p4</i>	64
17.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>Price</i>	66
18.	Listado de los modelos lineales simples elaborados para cada uno de los cuales se evalúa: coeficiente de determinación R^2 , R^2 Ajustado, el error <i>RMSE</i> y el p-value asociado al test de Wall	75

19.	Ranking de los 5 modelos lineales que ofrecen mejores resultados dadas las métricas evaluadas. Para cada modelo se vuelve a mostrar las métricas: coeficiente de determinación R^2 , R^2 Ajustado, el error $RMSE$ y el p-value asociado al test de Wall	77
20.	Ranking de los 5 modelos lineales que ofrecen mejores resultados dadas las métricas evaluadas. Para cada modelo se vuelve a mostrar las métricas: coeficiente de determinación R^2 , R^2 Ajustado, el error $RMSE$ y el error $RMSE$ medido mediante validación cruzada	78
21.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P27p4$ elevado al cuadrado	82
22.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P27p4$ elevado a 3	82
23.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P27p4$ elevado a 4	82
24.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P11p4$ elevado a 2	83
25.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H13p1$ elevado a 2 y eliminando el término con $P11p4$ al cuadrado	83
26.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H13p1$ elevado a 2	85
27.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H40p4$ elevado a 2	85
28.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P16p2$ elevado a 2	86
29.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P1$ elevado a 2	86
30.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P1$ elevado a 2	87
31.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P11p4$ elevado a 2	88
32.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P6p2$ elevado a 2	89
33.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H8p2$ elevado a 2	90
34.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P8p2$ elevado a 2	91
35.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H10p1$ elevado a 2	91

36.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H18pA$ elevado a 2	92
37.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H13p1$ elevado a 2	92
38.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P14p9$ elevado a 2	93
39.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con la raíz de $P12p2$	95
40.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con la raíz de $H18pA$	95
41.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con la raíz de $P15p3$	96
42.	Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con la raíz de $P1$	96
43.	Métricas de rendimiento evaluadas para el modelo compuesto que sustituye el término cuadrático con $H10p1$ por otro término elevado a 4.	97
44.	Métricas de rendimiento evaluadas para el modelo 7-NN elaborado con todas las variables del <i>dataset</i> con todos sus parámetros.	98
45.	Métricas de rendimiento evaluadas para el modelo 7-NN elaborado con todas las variables del <i>dataset</i> con todos sus parámetros y $P1$ escalada al intervalo [0,1].	98
46.	Métricas de rendimiento evaluadas para el modelo 7-NN elaborado con todas las variables del <i>dataset</i> a las cuales se han aplicado transformaciones para hacerlas tender a la normalidad y la variable $P1$ ha sido escalada al intervalo [0,1].	99
47.	Métricas de rendimiento evaluadas para el modelo 7-NN elaborado con todas las variables del <i>dataset</i> a las cuales se han aplicado transformaciones para hacerlas tender a la normalidad. En esta ocasión, no se ha reescalado $P1$	100
48.	Métricas de rendimiento evaluadas los modelos diferentes modelos generados para cada valor de k , a partir del modelo que aplicaba transformaciones a las variables del <i>dataset</i> para hacerlas tender a distribuciones normales. Se ha resaltado en negro las métricas óptimas.	100
49.	Métricas de rendimiento evaluadas para los diferentes modelos generados para cada valor de k , a partir del modelo que reescalaba la variable $P1$. Se ha resaltado en negro los valores óptimos.	101
50.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>Compactness</i>	108

51.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>Compactness</i>	112
52.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>Circularity</i>	. 112
53.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>Distance_circularity</i>	115
54.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>Radius_ratio</i>	. 118
55.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>Praxis_aspect_ratio</i>	122
56.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>Max_length_aspect_ratio</i>	124
57.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>Scatter_ratio</i>	127
58.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>Elongatedness</i>	128
59.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de <i>Kurtosis</i> calculados para el atributo <i>Praxis_rectangular</i>	130
60.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo <i>Length_rectangular</i>	132
61.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo <i>Major_variance</i>	134
62.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo <i>Minor_variance</i>	136
63.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo <i>Gyration_radius</i>	137
64.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo <i>Major_skewness</i>	139
65.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo <i>Minor_skewness</i>	141
66.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo <i>Minor_kurtosis</i>	143
67.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo <i>Major_kurtosis</i>	144
68.	Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo <i>Hollows_ratio</i>	146
69.	Tabla de contingencia con la distribución de la variable <i>Class</i>	. 147
70.	Métricas de rendimiento evaluadas para el modelo 7-NN elaborado con todas las variables del <i>dataset</i>	151

71.	Métricas de rendimiento evaluadas para el modelo 7-NN elaborado con todas las variables del <i>dataset</i> a las cuales se ha aplicado un escalado	152
72.	Métricas de rendimiento evaluadas para los diferentes modelos generados para cada valor de k , a partir del modelo que reescalaba todas las variables independientes. Se ha resaltado en negro los valores óptimos.	153
73.	Varianzas de todas las variables independientes del <i>dataset</i> calculadas por clase.	154
74.	Modelos LDA ignorando los atributos <i>Scatter_ratio</i> , <i>Praxis_rectangular</i> , <i>Length_rectangular</i> , <i>Minor_ariance</i> y <i>Gyration_radius</i>	155
75.	Modelos LDA desarrollados con todas las variables del <i>dataset</i>	156
76.	Modelos QDA desarrollados con todas las variables del <i>dataset</i>	157
77.	Comparación los mejores modelos obtenidos	158

1. Introducción

En el presente documento se recoge el informe elaborado como resultado de la ejecución del Trabajo integrador de la asignatura *Introducción a la Ciencia de Datos* consistente, en primer lugar, en la realización de diversos análisis exploratorios de datos sobre dos *datasets*, con el objetivo de **determinar información preliminar y relevante** sobre sus variables y sobre la distribución de los mismos y, en segundo lugar, en la **generación de diversos modelos óptimos de clasificación y regresión** (según corresponda) a partir de la información contenida en los *datasets* para la predicción de nuevos datos y la evaluación del rendimiento de los mismos.

Los *datasets* que se tratarán son **house**, el cual plantea un problema de regresión y sobre el cuál se tratarán de obtener diversos modelos óptimos de regresión lineal simples y múltiples, modelos de regresión basados en *k-NN* (*k Nearest Neighbour*) y otros modelos de regresión no lineales; y **vehicle**, para el cual, se plantea un problema de clasificación que se tratará de resolver mediante la construcción de modelos de clasificación basados en *k-NN*, modelos basados en *LDA* y modelos basados en *QDA*.

2. El dataset *house*: Problema de regresión.

2.1. Descripción del dataset *house*

El dataset *house* fue elaborada a partir del censo realizado en Estados Unidos en 1990 y recoge información sobre la composición demográfica y el estado del mercado inmobiliario con el objetivo de predecir el precio medio de las viviendas de cada región en función de esta información demográfica.¹

Este *dataset* consta de 16 variables independientes y una variable real dependiente, asimismo y dentro de él se recogen un total 22784 instancias que representan las diferentes regiones para las cuales se han medido los parámetros antes mencionados y el precio medio de las viviendas que se localizan en los mismos.

Las variables independientes son las siguientes²:

¹Enlace al repositorio donde se obtuvieron los datos:
<https://sci2s.ugr.es/keel/dataset.php?cod=95>

²Enlace a la documentación de la API con SF1:

- **P1:** Variable numérica entera que refleja la población total registrada en la región.
- **P5p1:** Variable numérica real que refleja el porcentaje de la población de sexo masculino.
- **P6p2:** Variable numérica real que refleja el porcentaje de la población de raza negra.
- **P11p4:** Variable numérica real que refleja el porcentaje de la población con 5 años.
- **P14p9:** Variable numérica real que refleja el porcentaje de mujeres viudas que forman parte de la población.
- **P15p1:** Variable numérica real que refleja el porcentaje de la población que vive en un hogar familiar.
- **P15p3:** Variable numérica real que refleja el porcentaje de la población que vive en residencias (incluídos presos en prisiones).
- **P16p2:** Variable numérica real que refleja el porcentaje de hogares de acogida con 2 o más personas y que constituyen un núcleo familiar.
- **P18p2:** Variable numérica real que refleja el porcentaje de hogares de acogida con más de una persona menor de 18 años y que no constituyen un núcleo familiar.
- **P27p4:** Variable numérica real que refleja la población de familias con 5 personas
- **H2p2:** Variable numérica real que refleja el porcentaje de unidades inmobiliarias vacías.
- **H8p2:** Variable numérica real que refleja el porcentaje de unidades inmobiliarias ocupadas cuyo arrendatario es de raza negra.
- **H10p1:** Variable numérica real que refleja el porcentaje de unidades inmobiliarias ocupadas cuyo arrendatario no es de origen hispano.
- **H13p1:** Variable numérica real que refleja el porcentaje de hogares entre 1 y 4 habitaciones

<https://api.census.gov/data/1990/sf1/variables.html>

- **H18pA:** Variable numérica real que refleja el porcentaje medio de las personas que ocupan hogares propios.
- **H40p4:** Variable numérica real que refleja el número de unidades inmobiliarias en venta cuyo precio de venta es inferior a 2 millones (mos)

Por su parte, la variable dependiente de este *dataset* es *Price* que constituye un valor numérico entero que representa precio asociado a cada unidad inmobiliaria.

2.2. Análisis exploratorio del dataset *house*

Una vez introducidos los atributos que constituyen este *dataset* y el significado de las instancias que contienen, se procede a realizar un **análisis exploratorio** de la información contenida en dicho **dataset** con el objetivo de estudiar distribución de sus atributos y determinar cualquier característica relevante que permita conocer más detalladamente el problema y que pueda resultar útil para la elaboración de modelos predictivos.

Primeramente, se analizará el *dataset* en busca de *Missing Values* (valores perdidos), para ello, se procede a ejecutar el siguiente trozo de código obteniendo la respuesta que se muestra debajo:

```
1 # Obtener información sobre Missing Values
2 any(is.na(house))
```

Script 1: Sentencia para comprobar la existencia de algún valor perdido en el *dataset*

```
[1] FALSE
```

De este modo encontramos que nuestro *dataset* carece de *Missing Values*.

2.2.1. Análisis de las variables del dataset

Para iniciar el proceso de análisis, se comenzará estudiando las medidas de posición y de dispersión de cada uno de los atributos del *dataset*, así como otras medidas de normalidad en las distribuciones de los atributos.

Como medidas de posición se evaluarán los valores mínimo y valores máximos de cada variable, la media aritmética , el primer cuartil, la mediana y el tercer cuartil. La dispersión de las distribuciones se evaluará gracias a la

desviación estándar, mientras que la normalidad se estudiará con los coeficientes de asimetría y de *Kurtosis*.

Todos estos estadísticos se computan de manera conjunta para cada variable mediante el siguiente conjunto de sentencias en R:

```
1 # Determinar los estadísticos de posición:  
2 # Valores mínimo y máximos, media, 1er cuartil, mediana,  
3     3er cuartil  
4 cat('Estadísticos de posición: Valores mínimo y máximos,  
5     media, 1er cuartil,',  
6     ' mediana, 3er cuartil', fill=T)  
7 summary(house)  
8  
9 # Determinar los estadísticos de dispersión: desviación  
10    típica  
11 cat('Desviación típica de los atributos', fill=T)  
12 apply(house, MARGIN=2, FUN=sd)  
13  
14 # Determinar coeficientes de Skew y Kurtosis  
15 cat('Coeficientes de Skew y Kurtosis', fill=T)  
16 skew_kurtosis <- apply(house, MARGIN=2, FUN=function(x)  
17     {c(skewness(x),  
18         kurtosis(x))})  
19 rownames(skew_kurtosis) <- c('Skew', 'kurtosis')  
20 skew_kurtosis
```

Script 2: Conjunto de sentencias para el calculo de los principales estadísticos de posición, dispersión y los coeficientes de asimetría y de Kurtosis de cada uno de los atributos del dataset

Se estudia cada variable con detalle:

- **P1:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	P1
Valor mínimo	2
Primer cuantil	427
Mediana	1346
Media	7809
Tercer cuantil	4518
Valor máximo	7322564
Desviación estándar	65872.42
Coeficiente de asimetría	72.74511
Coeficiente de Kurtosis	7202.97524

Cuadro 1: Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo P1

Los anteriores estadísticos nos muestran que la distribución de la variable $P1$ presenta una **pronunciada asimetría hacia la izquierda de la distribución**, hecho que se ve claramente reflejado si se compara el valor inferior de la Mediana, e incluso del tercer cuantil, respecto de la Media, lo que implica que un porcentaje superior al 75 % de la distribución se sitúa por debajo de la media.

Por su parte, el valor positivo del coeficiente de Kurtosis nos lleva a considerar una **distribución leptocúrtica** con una amplia dispersión de los datos, lo cual plantea la existencia de valores *outliers* situados a la derecha de la distribución, con valores muy superiores respecto de los valores centrales de la distribución.

Se decide estudiar este fenómeno de forma gráfica haciendo uso de un diagrama *boxplot* y del siguiente conjunto de sentencias para dibujarlo:

```

1 # Diagrama boxplot
2 graf <- ggplot(house, aes(x='P1', y=P1)) +
3   geom_boxplot(outlier.alpha=0.4, outlier.colour='
4   red', outlier.shape=8) +
5   labs(cation='center', y=' ', x=' ')
6 graf
7
8 # Diagrama boxplot en escala logarítmica
9 hraf <- ggplot(house, aes(x='P1', y=P1)) +
10  geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
11  outlier.shape=8) +
12  scale_y_log10() +
13  labs(cation='center', y=' ', x=' ')

```

```
12 hraf
```

Script 3: Sentencias para la elaboración de un diagrama *boxplot* de la variable $P1$

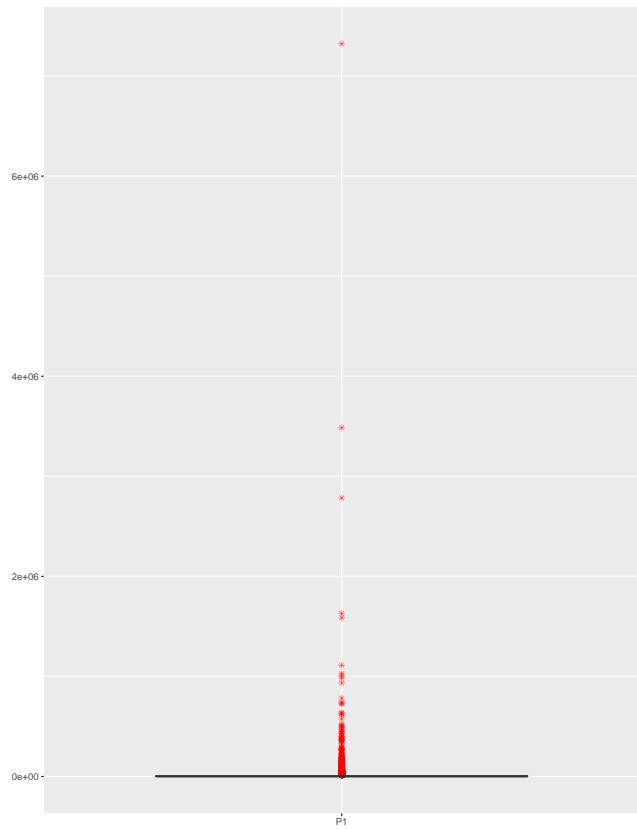


Figura 1: Diagrama *boxplot* de la variable $P1$ de *house*

En la anterior gráfica se refleja de una forma muy pronunciada, la concentración de la mayor parte de la distribución en una región muy reducida, presentando la distribución numerosos *outliers* con valores muy superiores y alejados de esta región.

La anterior gráfica no permite reflejar con claridad la forma de la distribución, por lo que se elabora otro diagrama *boxplot* en escala logarítmica:

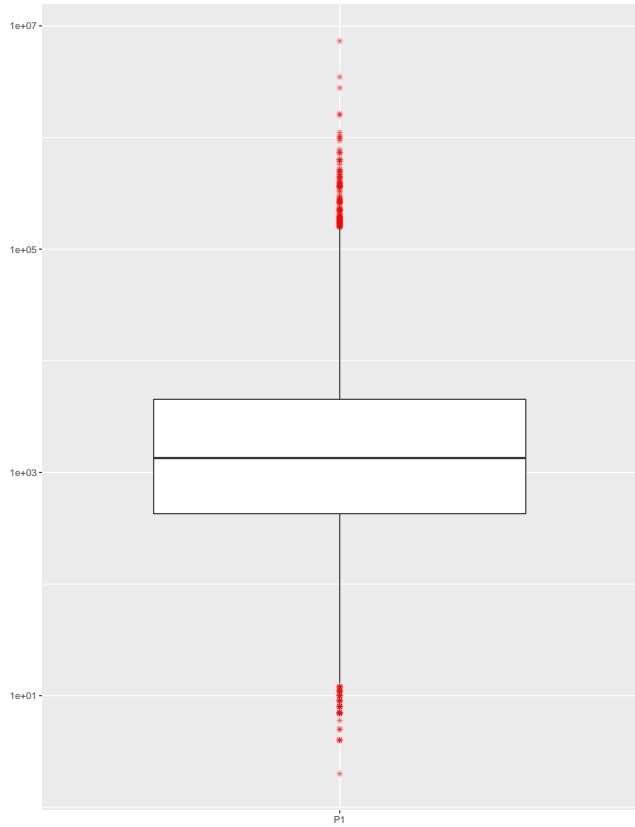


Figura 2: Diagrama *boxplot* de la variable P1 de *house* en escala logarítmica en base 10

Se decide estudiar más detenidamente y de forma separada, la distribución de los *outliers* y la del resto de la distribución:

Tomando como referencia el rango de distribución representado en el diagrama *boxplot*, es decir, aquellos valores comprendidos en el rango $[Q1 - 1,5 * IQR, Q3 + 1,5 * IQR]$, donde IQR es el rango intercuartílico y $Q1$ y $Q3$ son, respectivamente, el primer y tercer quartil; se divide el conjunto de datos en tres rangos: $(-\infty, Q1 - 1,5 * IQR)$, $[Q1 - 1,5 * IQR, Q3 + 1,5 * IQR]$ y $(Q3 + 1,5 * IQR, +\infty)$ y haciendo uso del siguiente conjunto de instrucciones que nos proporciona la salida que se muestra a continuación:

```

1 # Aislar la distribución comprendida entre [Q1-1.5*IQR ,
  Q3+1.5*IQR]
2 quantiles.P1 <- quantile(house$P1)
3 iqr.P1 <- IQR(house$P1)

```

```

4
5 centro.house.P1 <- house %>% select(P1) %>%
6   filter(P1 > (quantiles.P1[2]-1.5*iqr.P1) & (P1 <
7     quantiles.P1[4]*1.5))
8
9 # Obtener estadísticos de posición y número de medidas
10 cat('Rango [Q1-1.5*IQR, Q3+1.5*IQR]:', fill=T)
11 centro.house.P1 %>% summarise(min=min(centro.house.P1$P1
12   ),
13   q1=quantile(centro.
14   house.P1$P1)[2],
15   median=median(centro.
16   house.P1$P1),
17   mean=mean(centro.
18   house.P1$P1)[4],
19   max=max(centro.house
20   .P1$P1),
21   count=n(),
22   ratio=n()/length(
23   house$P1))

```

Script 4: Conjunto de sentencias para dividir el conjunto de valores de P1 en 3 intervalos y calcular medidas de posición: Valor mínimo, valor máximo, primer cuartil, mediana, tercer cuartil y media aritmética, el número de valores incluídos en cada intervalo así como el ratio de la distribución de P1 que representan

```
Rango [Q1-1.5*IQR, Q3+1.5*IQR]:
  min   q1 median      mean    q3   max count      ratio
1    2 329     908 1525.083 2179 6777 18523 0.8129828
```

```
Rango (-Inf, Q1-1.5*IQR):
```

```
Error: Evaluation error: only defined on a data frame with all numeric variab
```

```
Rango (Q3+1.5*IQR, +Inf):
```

```
  min   q1 median      mean    q3   max count      ratio
1 10656 14562 21997.5 47401 39780.25 7322564 2918 0.1280723
```

La ejecución del anterior conjunto de sentencias nos devuelve error para el intervalo $(-\infty, Q1 - 1,5 * IQR)$ puesto que los *outliers* que se incluían aquí, han sido incluídos en el intervalo donde se concentra la mayor parte de la distribución ($[Q1 - 1,5 * IQR, Q3 + 1,5 * IQR]$) por

su proximidad numérica.

Asimismo, se observa que el intervalo $[Q1 - 1,5 * IQR, Q3 + 1,5 * IQR]$ supone un porcentaje superior al 81 % de la distribución original, mientras que los *outliers* extremos se aproximan al 13 %, con lo cual, **se podría optar por eliminarlos en caso de que nos dificultara el análisis o la posterior obtención de modelos.**

Por último, se decide representar mediante histogramas los valores del intervalo $[Q1 - 1,5 * IQR, Q3 + 1,5 * IQR]$ y los valores del intervalo $(Q3 + 1,5 * IQR, +\infty)$ con la finalidad de conocer con detalle la forma de la distribución.

```
1 # Obtener los outliers inferiores a Q1-1.5*IQR
2 cat('Rango (-Inf, Q1-1.5*IQR):', fill=T)
3 low.outliers.house.P1 <- house %>% select(P1) %>%
4   filter(P1 < (quantiles.P1[2]-1.5*iqr.P1))
```

Script 5: Conjunto de sentencias para dibujar y visualizar el histograma del intervalo donde se recoge casi la totalidad de la distribución

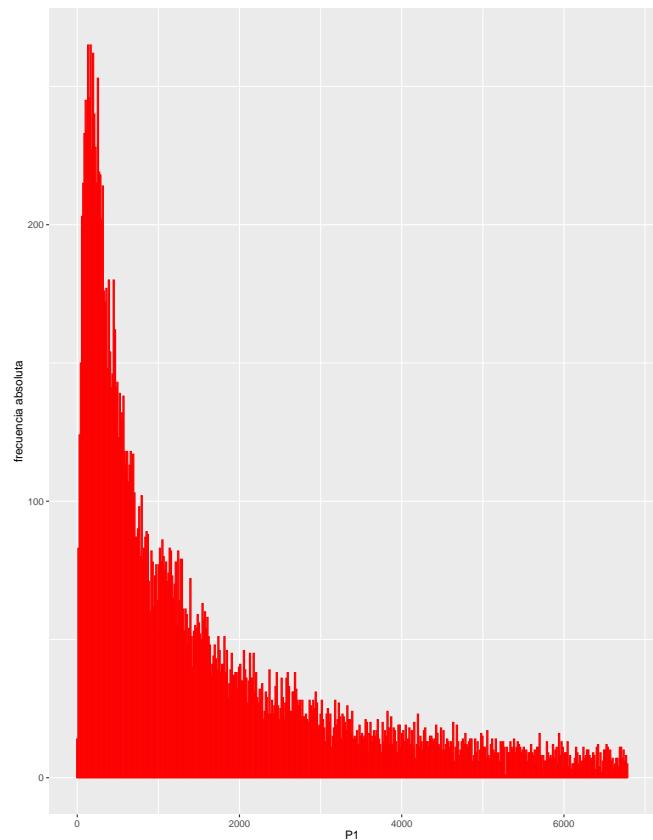


Figura 3: Histograma del subconjunto de la variable $P1$ comprendida en el intervalo $[Q1 - 1,5 * IQR, Q3 + 1,5 * IQR]$

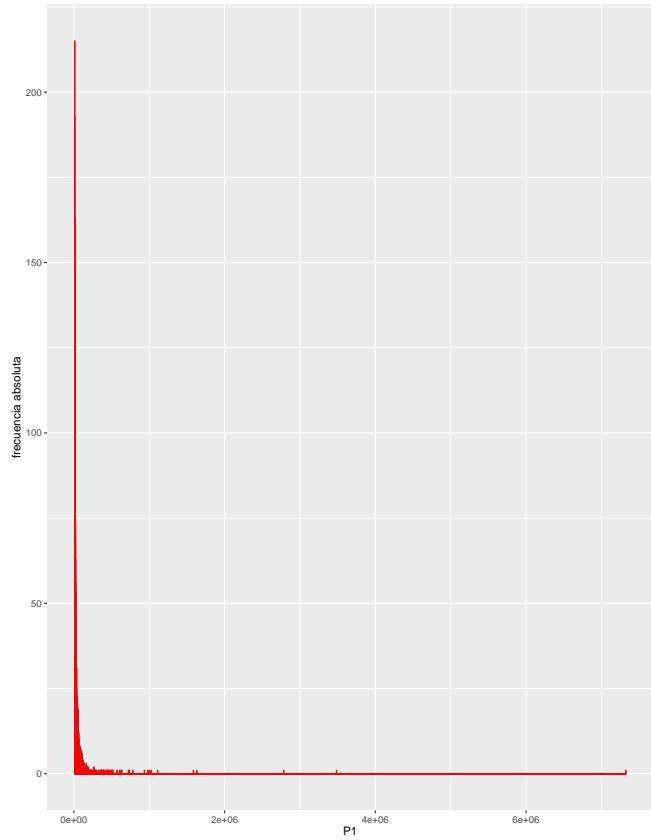


Figura 4: Histograma del subconjunto de la variable $P1$ comprendida en el intervalo $(Q3 + 1,5 * IQR, +\infty)$

Considerando ambos histogramas, se observa que la distribución conforme aumenta el valor de $P1$ sufre un incremento brusco de su densidad, hasta alcanzar su máxima densidad, tras lo cual la densidad de la distribución decrece de forma exponencial hasta alcanzar el valor máximo de la distribución. Por ello, se considera que **la forma de la distribución de esta variable podría aproximarse por una distribución beta**.

- **P5p1:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	P5p1
Valor mínimo	0.1250
Primer cuantil	0.4646
Mediana	0.4804
Media	0.4823
Tercer cuantil	0.4960
Valor máximo	0.9231
Desviación estándar	0.03187648
Coeficiente de asimetría	2.128916
Coeficiente de Kurtosis	21.878934

Cuadro 2: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* caclulados para el atributo *P5p1*

Los anteriores estadísticos nos muestran que la distribución se encuentra desplazada hacia la izquierda y que presenta una amplia dispersión respecto al centro de densidad de la distribución, lo cual plantea nuevamente la existencia de *outliers*.

Para analizar de forma gráfica la distribución de esta variable, elaboramos un diagrama *boxplot*:

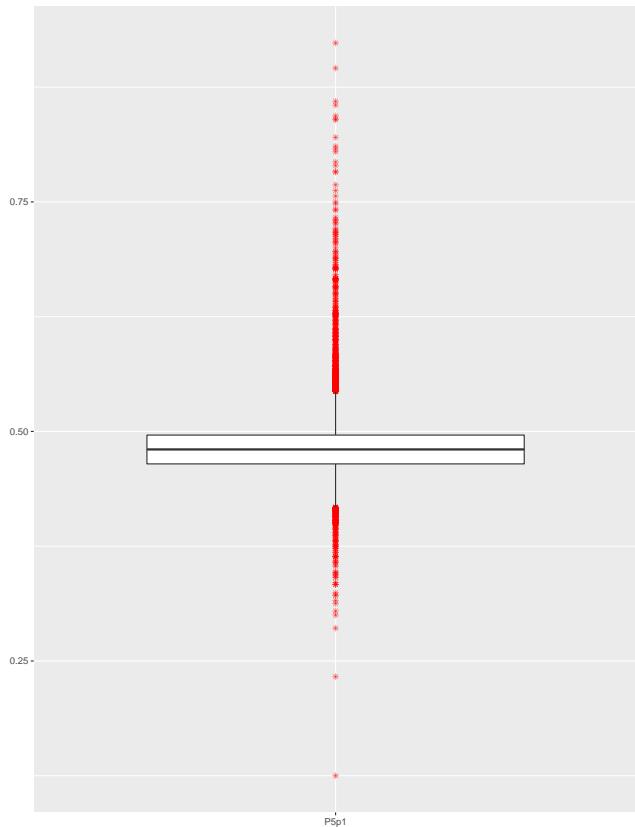


Figura 5: Diagrama *boxplot* de la variable *P5p1*

El anterior diagrama nos muestra que, efectivamente la distribución presenta su región más densa en su centro, la cual es estrecha respecto del intervalo de valores para esta variable y presenta una amplia dispersión respecto de su centro con menor densidad.

Finalmente, se dibuja un histograma para analizar la forma de la distribución:

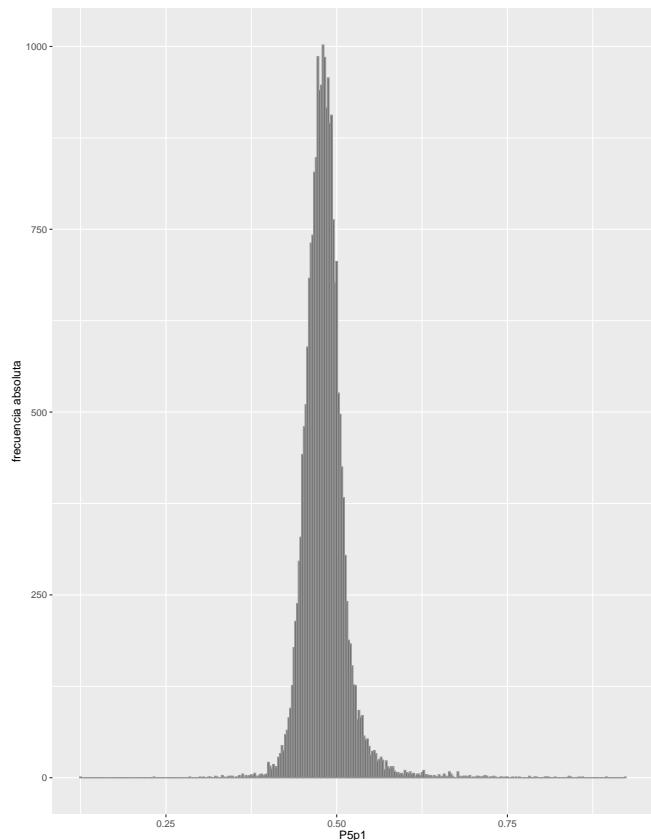


Figura 6: Histograma de la variable $P5p1$

Se observa por tanto que **la distribución es leptocúrtica con el centro algo desplazado hacia la izquierda**.

Interpretando estos resultados, observamos que en la mayor parte de las regiones recogidas en el *dataset* el porcentaje de la población de sexo masculino tiende a 0.5, es decir, **en la mayor parte de las regiones, el número de hombres y mujeres tiende a ser casi idéntico, aunque por lo general, el porcentaje de hombres tienda a ser algo inferior**.

- **P6p2:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	P6p2
Valor mínimo	0
Primer cuantil	0
Mediana	0.003413
Media	0.063982
Tercer cuantil	0.033377
Valor máximo	1
Desviación estándar	0.1509207
Coeficiente de asimetría	3.341087
Coeficiente de Kurtosis	15.168186

Cuadro 3: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* caclulados para el atributo *P6p2*

Para esta variable, se observa que la distribución presente una profunda asimetría hacia la izquierda de la distribución, y que presenta una amplia dispersión respecto de su centro de distribución, además al ser la media de la distribución superior a la mediana y al tercer cuartil, se podría pensar que la distribución presenta *outliers* a la derecha de la distribución.

Para analizar gráficamente la distribución y organización de esta variable representamos un diagrama *boxplot* y el histograma de la misma:

El diagrama *boxplot* originado es el siguiente:

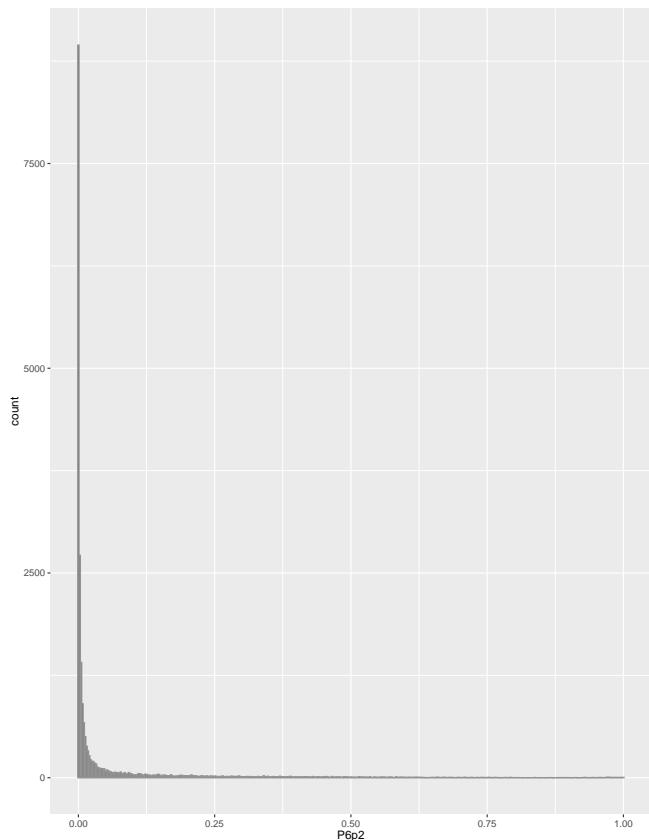


Figura 7: Diagrama *boxplot* de la variable $P6p2$

Mientras que el histograma originado es el siguiente:

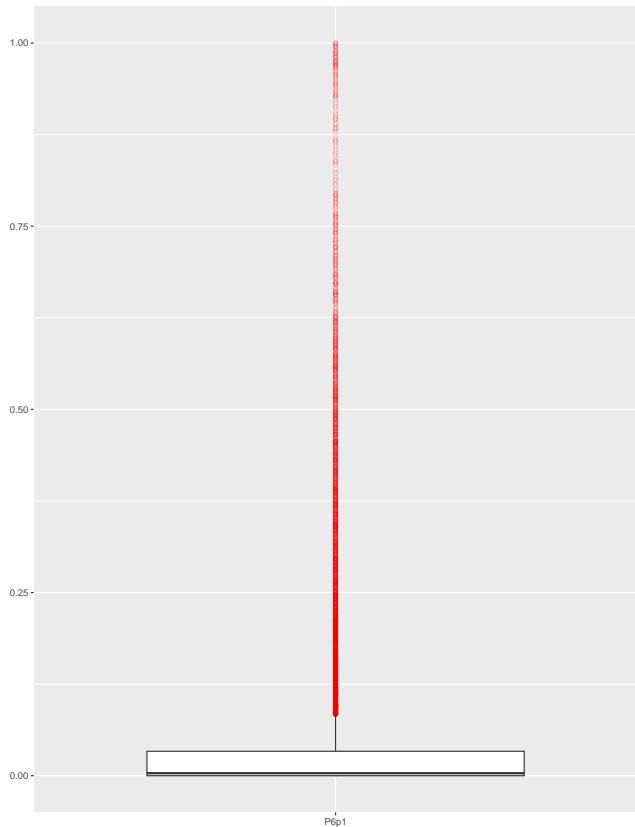


Figura 8: Histograma de la variable *P6p2*

Observando los anteriores gráficos, podemos observar que la distribución concentra su mayor densidad en una región muy estrecha y concentrada próxima al valor 0 y que conforme aumenta el valor de esta variable, la densidad sufre una disminución exponencial muy acentuada.

Esta distribución al igual que la de la variable *P1*, podría aproximarse también por una distribución beta o por una distribución basada en una función exponencial decreciente.

Interpretando estos resultados, podemos obtener un dato muy significativo: **La mayor parte de las regiones evaluadas en el dataset, presentan un porcentaje de población de raza negra muy reducido.**

- **P11p4:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	P11p4
Valor mínimo	0
Primer cuantil	0.1121
Mediana	0.1555
Media	0.1639
Tercer cuantil	0.2036
Valor máximo	0.91723
Desviación estándar	0.08046111
Coeficiente de asimetría	1.571398
Coeficiente de Kurtosis	10.673767

Cuadro 4: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* caclulados para el atributo *P11p4*

Los anteriores estadísticos nos dan una referencia de una distribución que se encuentra algo desplazada a la izquierda y que presenta un centro de distribución muy denso y concentrado en una región muy estrecha. El coeficiente de Kurtosis revela una amplia dispersión de los datos respecto a este centro de distribución, aunque en comparación con las anteriores variables, y obervando el valor de la media más próximo a la mediana que al tercer cuantil, se espera la existencia de menos *outliers* y situados a la derecha de la distribución, y/o se encuentran situados a la derecha de la distribución próximos al centro de distribución.

Nuevamente analizamos la distribución de la variable, así como su forma con ayuda de un diagrama *boxplot* y un histograma:

El diagrama *boxplot* originado es el siguiente:

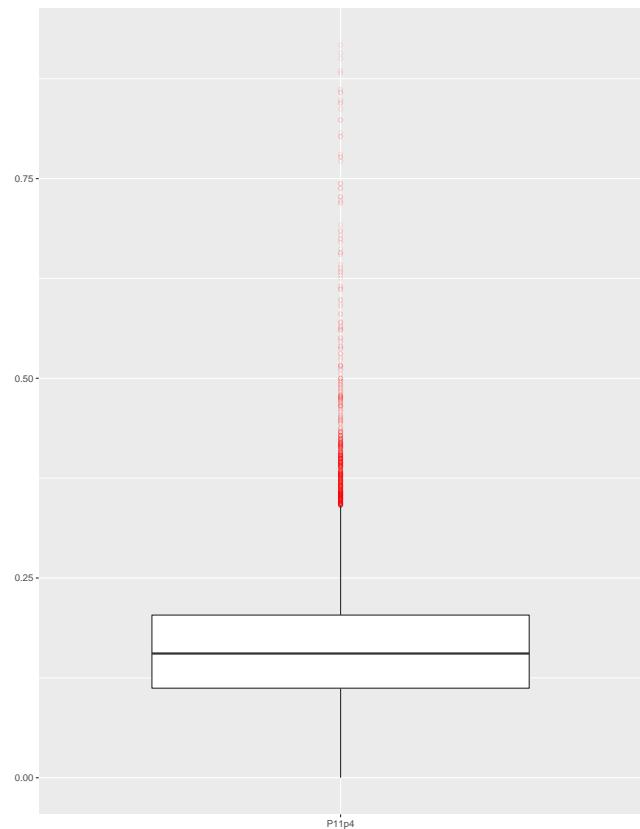


Figura 9: Histograma de la variable $P11p4$

Por su parte, la forma de la distribución se puede apreciar en el siguiente histograma:

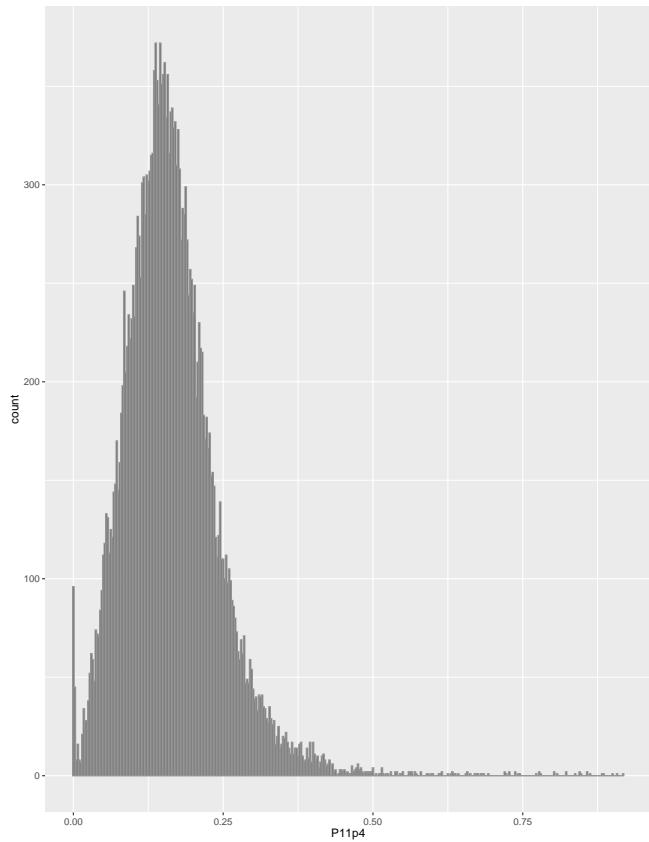


Figura 10: Histograma de la variable *P11p4*

Los anteriores gráficos nos permiten confirmar las conclusiones tomadas a partir de la información de los estadísticos.

Esta información nos revela que en la mayor parte de las regiones recogidas en el *dataset* el porcentaje de la población con 5 años se sitúa aproximadamente en torno al 10 % y 20 %. Por su parte, en el anterior histograma podemos observar la existencia de pequeño porcentaje de regiones donde el procentaje de niños y niñas de 5 años tiende a ser nulo.

- **P14p9:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	P14p9
Valor mínimo	0
Primer cuantil	0.08124
Mediana	0.11713
Media	0.12154
Tercer cuantil	0.15676
Valor máximo	0.051187
Desviación estándar	0.0566866
Coeficiente de asimetría	0.6611913
Coeficiente de Kurtosis	4.4649892

Cuadro 5: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* caclulados para el atributo *P14p9*

Los estadísticos nos dan una idea de una distribución ligeramente desplazada a la izquierda que presenta una región central densa y con dispersión de los datos respecto de esta región central, lo cual plantea nuevamente la existencia de *outliers* situados a la derecha de la distribución.

Para analizar con detalle este hecho, nuevamente volvemos a realizar un diagrama *boxplot* y un histograma de la distribución de esta variable:

Se muestra el diagrama *boxplot*:

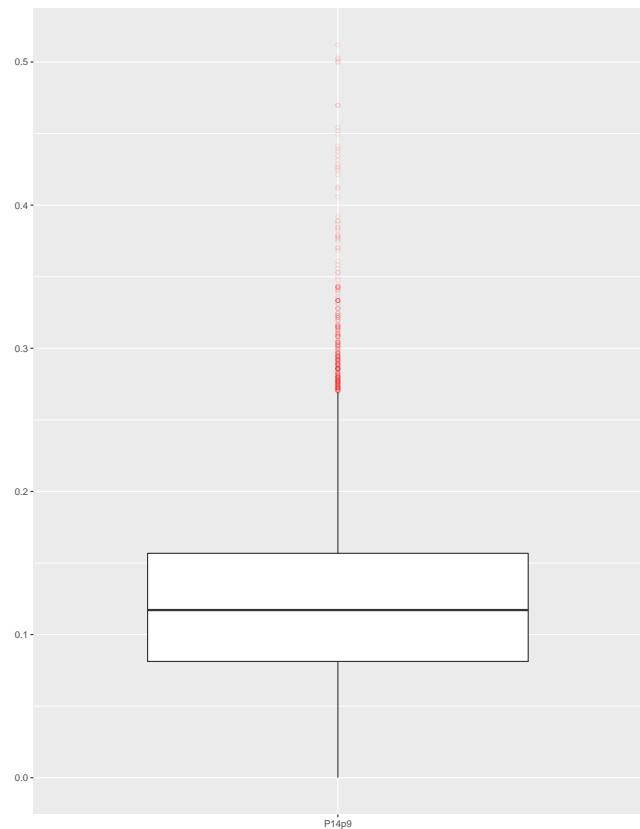


Figura 11: Histograma de la variable $P14p9$

Y el histograma con la forma de la distribución:

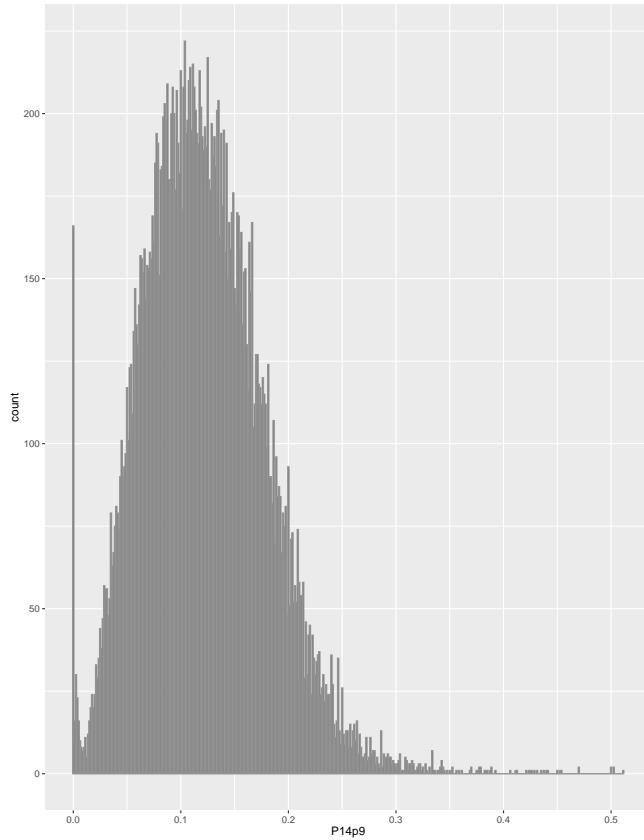


Figura 12: Histograma de la variable $P14p9$

Estos gráficos nos permiten confirmar las conclusiones formuladas. Por su parte, el histograma nos muestra una pequeña región densa muy próxima al valor 0.

Interpretando esta información, se observa que, en la mayor parte de las regiones evaluadas, el porcentaje de mujeres viudas respecto del resto de la población se sitúa aproximadamente entre el 8% y el 16% con regiones en los que este porcentaje tiende a ser nulo.

- **P15p1:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	P15p1
Valor mínimo	0.05416
Primer cuantil	0.81969
Mediana	0.86365
Media	0.85108
Tercer cuantil	0.89966
Valor máximo	1
Desviación estándar	0.07945483
Coeficiente de asimetría	-2.598684
Coeficiente de Kurtosis	16.436190

Cuadro 6: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* caclulados para el atributo *P15p1*

Los anteriores estadísticos nos llevan a considerar la distribución de la variable *P15p1* desplazada a la derecha con un centro de distribución situado en una región estrecha densa y una gran dispersión de los datos respecto de este centro de distribución, planteando la existencia de *outliers* situados a la izquierda de la distribución.

Se estudia esta distribución, así como su forma más detalladamente haciendo uso de un diagrama *boxplot* y un histrograma:

Se muestra el diagrama *boxplot*:

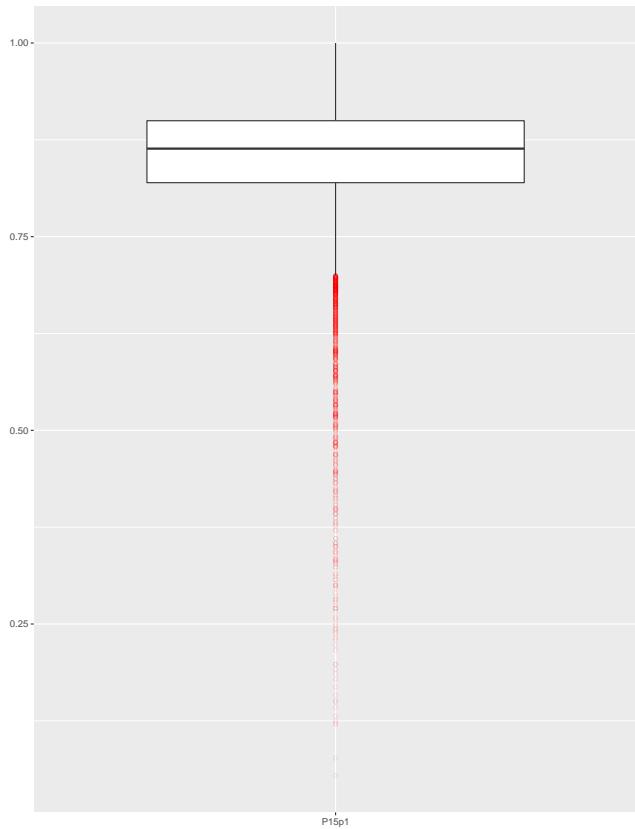


Figura 13: Histograma de la variable $P15p1$

Y el histograma con la forma de la distribución:

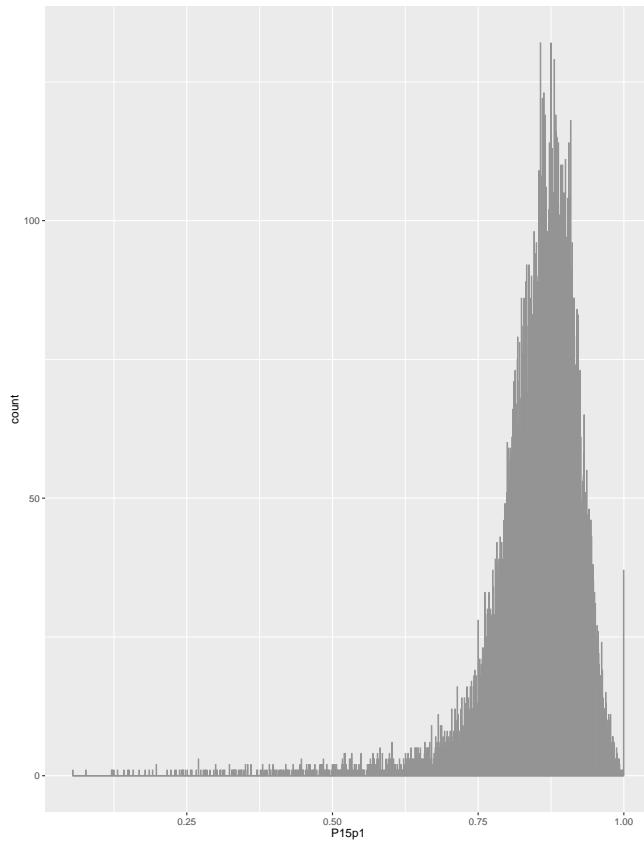


Figura 14: Histograma de la variable $P15p1$

Las anteriores gráficas nos permite verificar de forma clara estas conclusiones. Nuevamente, en histograma revela una pequeña región densa distinta al centro de la distribución, la cual se localiza para valores próximos a 1.

- **P15p3:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	P15p3
Valor mínimo	0
Primer cuantil	0
Mediana	0
Media	0.02049
Tercer cuantil	0.01920
Valor máximo	0.94332
Desviación estándar	0.0576660
Coeficiente de asimetría	6.729615
Coeficiente de Kurtosis	64.674322

Cuadro 7: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* caclulados para el atributo *P15p3*

Los estadísticos nos muestran una distribución con un centro muy denso y estrecho al valor 0. La distribución se halla, por tanto, fuertemente desplazada a la derecha y con dispersión de los datos respecto de su centro, lo cual plantea la existencia de *outliers*.

Se analiza detalladamente la distribución mediante un gráfico *boxplot* y la forma de la misma mediante un histograma:

Se muestra el diagrama *boxplot*:

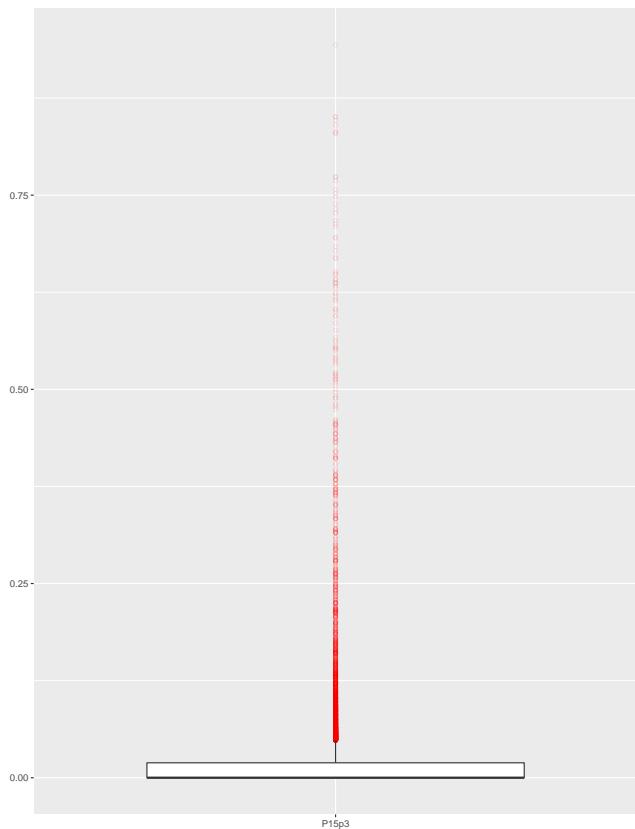


Figura 15: Histograma de la variable $P15p3$

Y el histograma con la forma de la distribución:



Figura 16: Histograma de la variable $P15p3$

Interpretando estos resultados, podemos observar que sólo en unas pocas regiones evaluadas en el *dataset* existen personas que viven en hogares con niños adoptados.

- **P16p2:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	P16p2
Valor mínimo	0.2337
Primer cuantil	0.6623
Mediana	0.7142
Media	0.7161
Tercer cuantil	0.7710
Valor máximo	1
Desviación estándar	0.08726448
Coeficiente de asimetría	-0.1347494
Coeficiente de Kurtosis	4.0850180

Cuadro 8: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *P16p2*

También se elabora a continuación un diagrama *boxplot* y un histograma para analizar de forma gráfica esta variable:

Se muestra el diagrama *boxplot*:

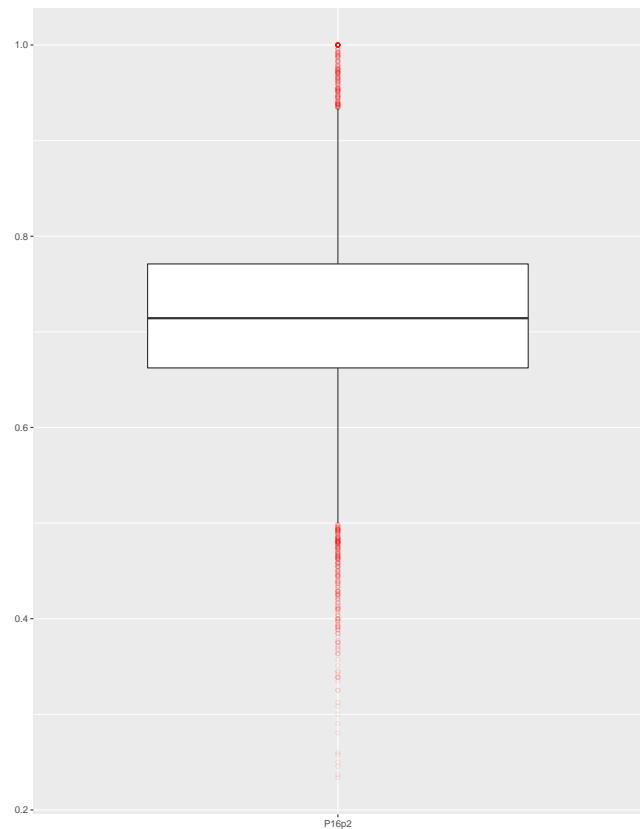


Figura 17: Histograma de la variable $P16p2$

Y el histograma con la forma de la distribución:

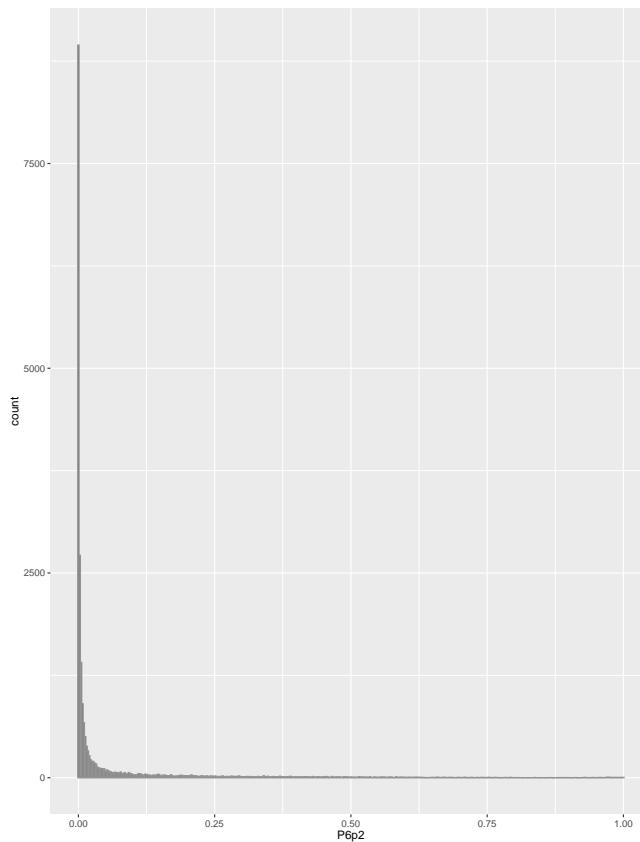


Figura 18: Histograma de la variable $P6p2$

A partir de los estadísticos anteriormente calculados y las anteriores gráficas, se determina que la distribución de esta variable presenta su centro ligeramente desplazado a la derecha, por su parte, la distribución presenta cierta dispersión de los datos respecto de su centro (distribución leptocúrtica), presentando, como se puede apreciar en el diagrama *boxplot*, *outliers* a ambos lados del centro de distribución.

- **P18p2:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	P18p2
Valor mínimo	0
Primer cuantil	0
Mediana	0.02591
Media	0.003589
Tercer cuantil	0.05007
Valor máximo	0.125000
Desviación estándar	0.005100383
Coeficiente de asimetría	5.377407
Coeficiente de Kurtosis	62.242068

Cuadro 9: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *P18p2*

Las anteriores medidas nos muestran una distribución fuertemente desplazada a la izquierda con un centro de distribución denso y concentrado en una región muy estrecha de la distribución. El coeficiente de kurtosis nos lleva a considerar que existe dispersión de los datos respecto del centro dando lugar a la aparición de *outliers* a ambos lados de la distribución.

Se representa, asimismo, el diagrama *boxplot* y el histograma de esta distribución:

```

1 graf
2
3 # Histograma
4 graf <- ggplot(house, aes(P18p2)) +
5   geom_histogram(binwidth=0.00125, color='gray58')
6 labs(y='frecuencia absoluta')
7 graf
8
9 # P18p2
10
11 # Diagrama boxplot
12 graf <- ggplot(house, aes(x='P18p2', y=P18p2)) +
13   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
14     outlier.shape=1) +

```

Script 6: Conjunto de sentencias para dibujar y visualizar un diagrama *boxplot* y un histograma de la variable *P18p2*

Se muestra el diagrama *boxplot*:

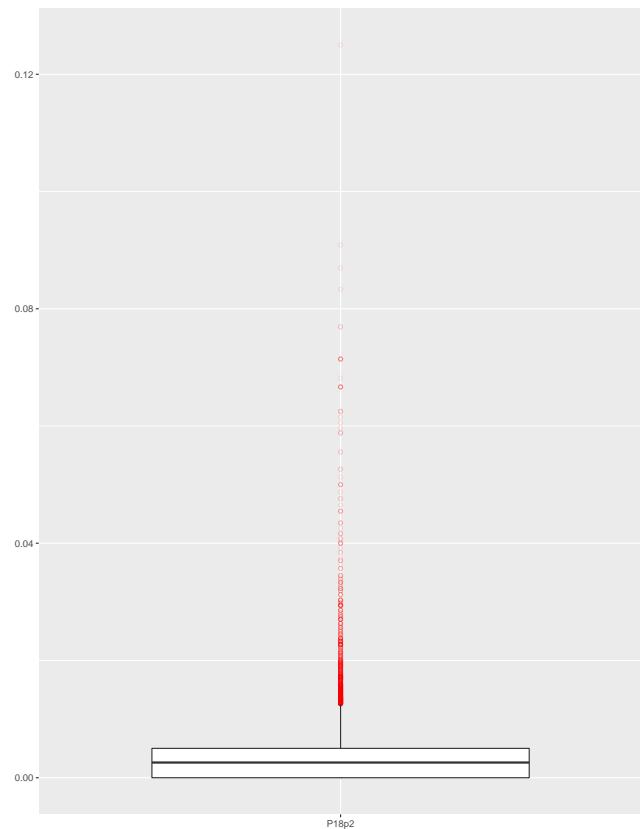


Figura 19: Histograma de la variable $P18p2$

Y el histograma con la forma de la distribución:

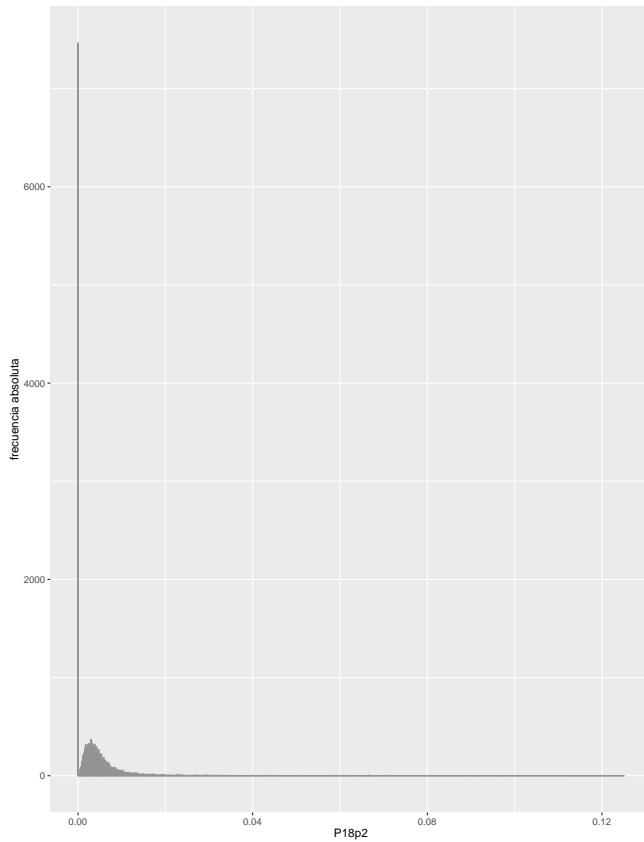


Figura 20: Histograma de la variable $P18p2$

Conviene aquí analizar en detalle el histograma representado, el cual muestra un aspecto de la distribución no fácilmente determinable con los estadísticos calculados anteriormente: en él se puede apreciar una zona de densidad mayor en torno al valor 0, tras lo cual aparece otra pequeña región densa en torno a la mediana de la distribución.

Se decide analizar en detalle esta región más densa, para ello, se decide estudiar y representar la región de la distribución definida en el intervalo $[Q1 - 1,5 * IQR, Q3 + 1,5 * IQR]$:

```

1 # Histograma del centro de la distribución de P18p2
2 iqr.P18p2 <- IQR(house$P18p2)
3 quantiles.P18p2 <- quantile(house$P18p2)
4
5 graf <- ggplot(house, aes(P18p2)) +
6   geom_histogram(binwidth=0.0001, color='gray58') +
7   labs(y='frecuencia absoluta') +

```

```

8   xlim(quantiles.P18p2[2]-1.5*iqr.P18p2, quantiles.P18p2
      [4]+1.5*iqr.P18p2)
9 graf

```

Script 7: Conjunto de sentencias para dibujar y visualizar un histograma de la región más densa de la variable $P18p2$

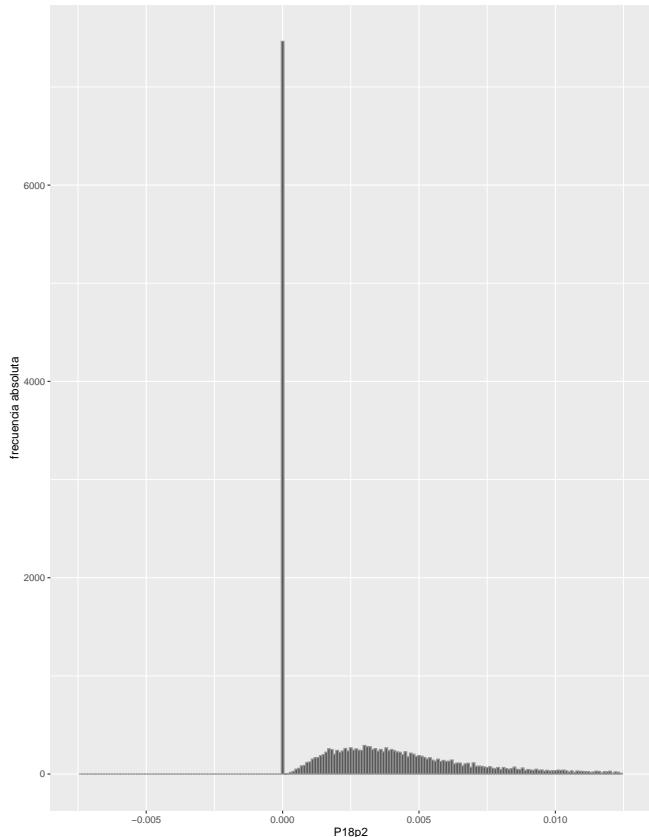


Figura 21: Histograma de la variable $P18p2$

En este histograma podemos observar más claramente este fenómeno: la región de mayor densidad de la distribución se halla en el valor 0, la cual constituiría de este modo la moda de la distribución de esta variable. Junto a esta región, aparece otra zona densa (mucho menos densa que la anterior) en torno a la mediana de la distribución. Esta característica dificulta en un principio, la aproximación de estos datos por una distribución conocida.

- **P27p4:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	P27p4
Valor mínimo	0
Primer cuantil	0.01626
Mediana	0.02752
Media	0.03326
Tercer cuantil	0.04283
Valor máximo	0.70574
Desviación estándar	0.02956722
Coeficiente de asimetría	4.063438
Coeficiente de Kurtosis	42.789326

Cuadro 10: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *P27p4*

Con estas medidas, podemos observar que la distribución nuevamente presenta una fuerte asimetría hacia la izquierda de la distribución, donde se localiza en una región estrecha el centro de la misma. Por último, la distribución presenta una amplia dispersión respecto de este centro de distribución, lo cual podría llevar a la aparición de *outliers* a la derecha de la distribución.

Nuevamente analizamos esta información de forma gráfica con ayuda de un diagrama *boxplot* y un histograma:

Se muestra el diagrama *boxplot*:

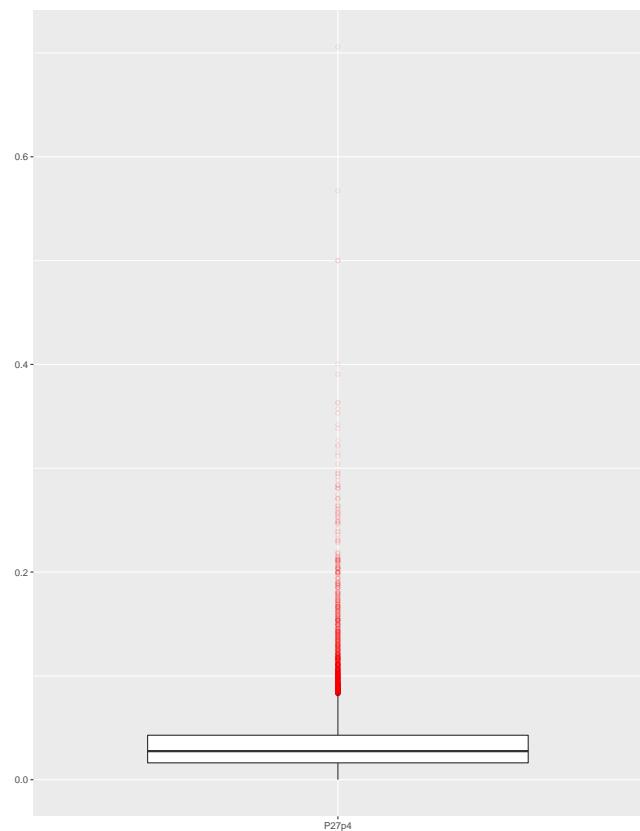


Figura 22: Histograma de la variable $P27p4$

Y el histograma con la forma de la distribución:

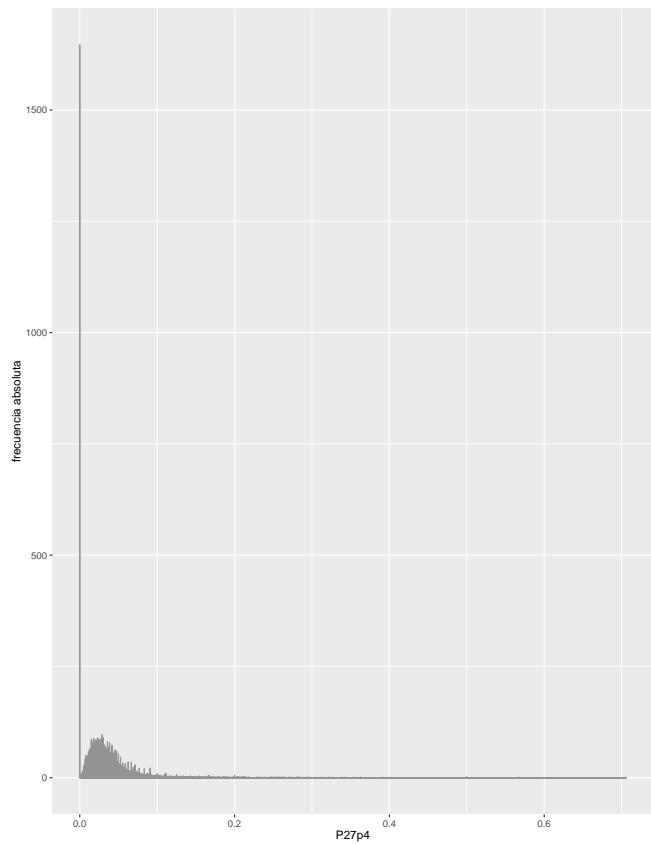


Figura 23: Histograma de la variable $P27p4$

Si se estudia el histograma de la distribución, observamos un fenómeno similar al que se determinó en la variable $P18p4$: La zona de mayor densidad de la distribución se localiza en el valor 0, apareciendo otra región densa en torno a la mediana de la distribución.

- **H2p2:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	H2p2
Valor mínimo	0
Primer cuantil	0.04930
Mediana	0.08118
Media	0.11053
Tercer cuantil	0.13143
Valor máximo	0.97518
Desviación estándar	0.1059248
Coeficiente de asimetría	3.18994
Coeficiente de Kurtosis	17.16255

Cuadro 11: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *H2p2*

La distribución de la variable *H2p2* presenta una asimetría hacia la izquierda de la distribución. El centro de la misma se halla localizado en una región estrecha, por su parte, el coeficiente de Kurtosis nos parece indicar una amplia dispersión de datos respecto de este centro de distribución, por lo que nuevamente podría dar lugar a la aparición de *outliers*.

Nuevamente analizamos esta información de forma gráfica con ayuda de un diagrama *boxplot* y un histograma:

Se muestra el diagrama *boxplot*:

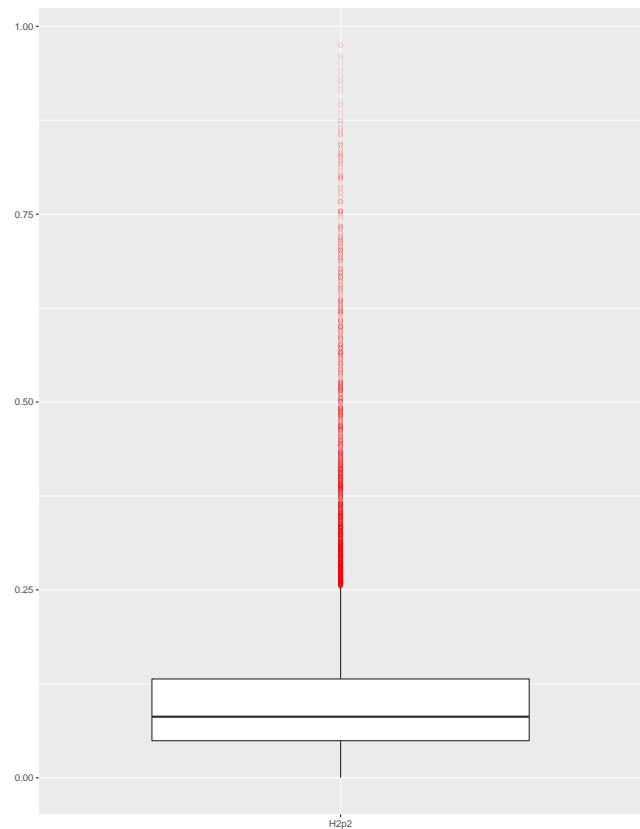


Figura 24: Histograma de la variable $H2p2$

Y el histograma con la forma de la distribución:

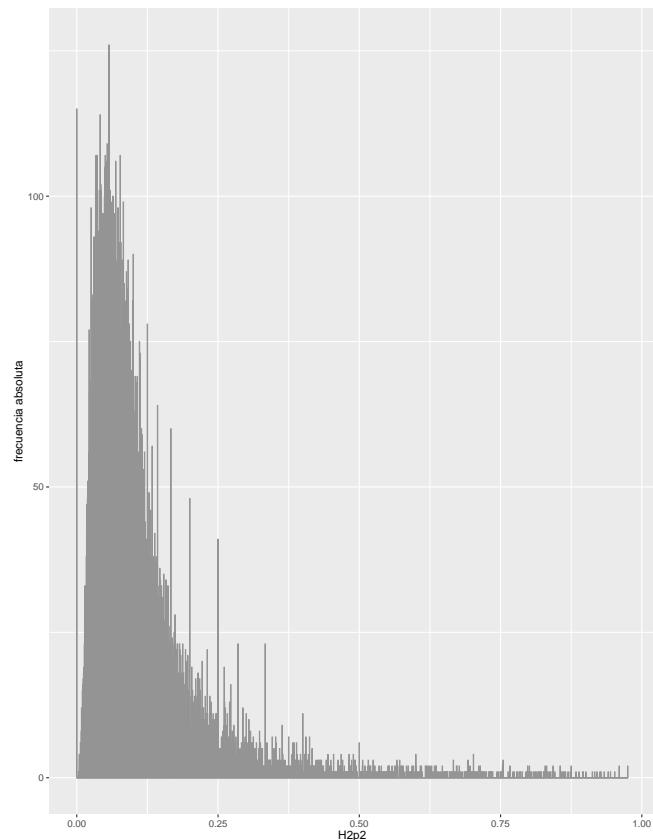


Figura 25: Histograma de la variable $H2p2$

Observando el histograma podemos apreciar como además del centro de la distribución localizado aproximadamente en torno a los valores 0.004 y 0.13, existe un buen porcentaje de regiones recogidas en el *dataset* para las cuales la variable adopta el valor 0.

- **H8p2:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	H8p2
Valor mínimo	0
Primer cuantil	0
Mediana	0.002538
Media	0.057437
Tercer cuantil	0.029928
Valor máximo	1
Desviación estándar	0.1398114
Coeficiente de asimetría	3.605403
Coeficiente de Kurtosis	17.835897

Cuadro 12: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *H8p2*

La distribución de la variable *H8p2* presenta un centro de distribución denso localizado en una región muy estrecha a la izquierda de la distribución con una amplia dispersión de datos respecto de este centro de distribución, tal y como indica el coeficiente de *Kurtosis*, lo que podría indicar la existencia de *outliers* a la derecha de la distribución. Nuevamente analizamos esta información de forma gráfica con ayuda de un diagrama *boxplot* y un histograma:

Se muestra el diagrama *boxplot*:

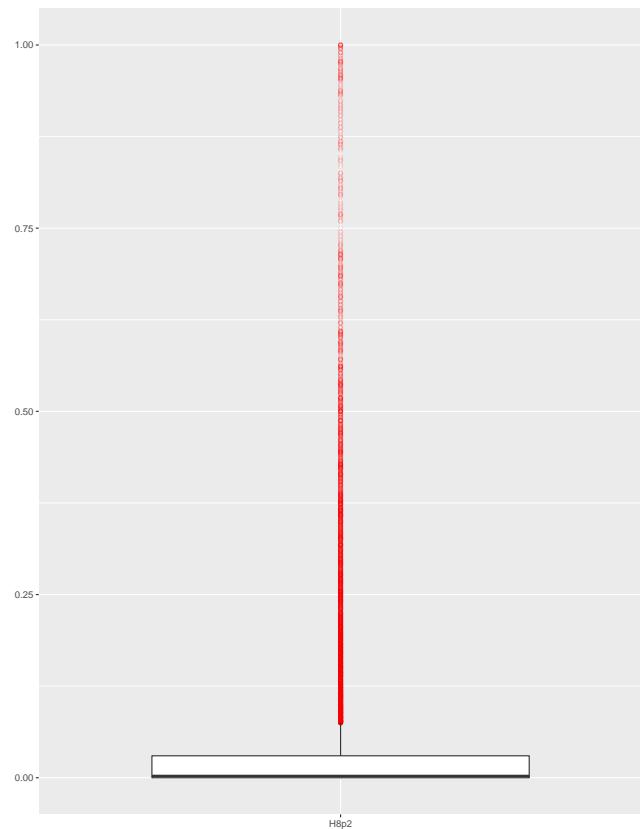


Figura 26: Histograma de la variable $H8p2$

Y el histograma con la forma de la distribución:

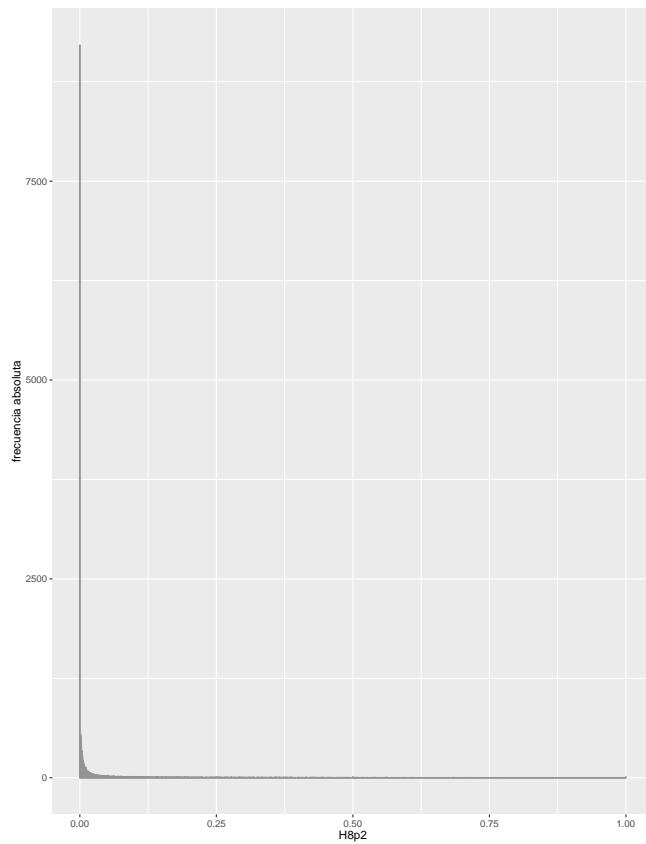


Figura 27: Histograma de la variable $H8p2$

En este caso, podemos comprobar como la distribución alcanza su densidad más alta para un valor de la variable igual a 0, y al incrementarse el valor de la variable, la densidad de la distribución decrece de forma exponencial siendo este decrecimiento muy acentuado.

- **H10p1:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	H10p1
Valor mínimo	0.003257
Primer cuantil	0.981410
Mediana	0.994100
Media	0.967045
Tercer cuantil	1
Valor máximo	1
Desviación estándar	0.09995253
Coeficiente de asimetría	-5.87264
Coeficiente de Kurtosis	42.87955

Cuadro 13: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *H10p1*

En esta ocasión, la distribución de la variable *H10p2* presenta un centro de distribución denso localizado en una región muy estrecha a la derecha de la distribución (muy próxima al valor máximo). El coeficiente de *Kurtosis* nos indica la existencia de una amplia dispersión de datos respecto de este centro de distribución, lo cual plantea la presencia de *outliers* a la izquierda de la distribución.

Como en todos los casos, se analiza esta información de forma gráfica con ayuda de un diagrama *boxplot* y un histograma:

Se muestra el diagrama *boxplot*:

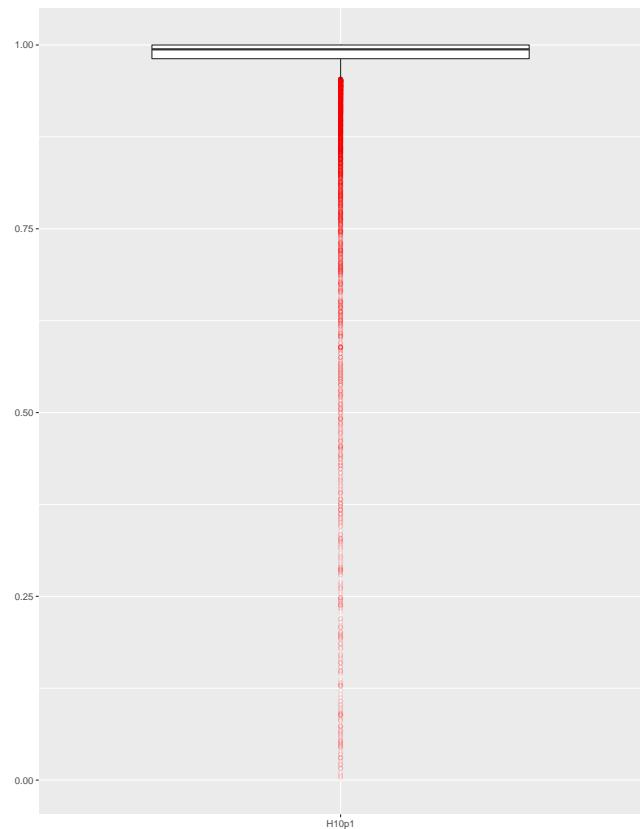


Figura 28: Histograma de la variable $H10p1$

Y el histograma con la forma de la distribución:

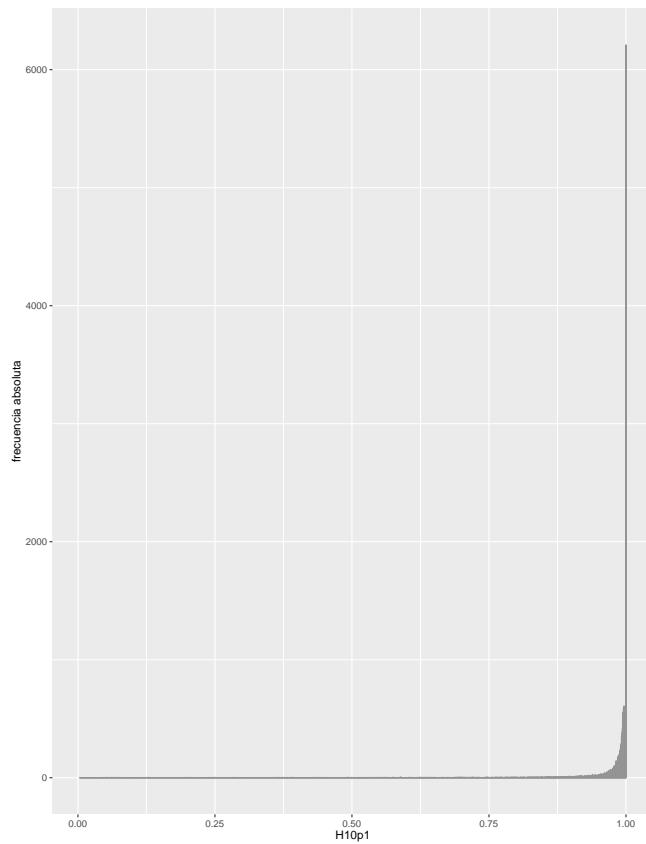


Figura 29: Histograma de la variable $H10p1$

Observamos pues que el centro de la distribución se halla estrechamente concentrado en el valor máximo de la distribución.

- **H13p1:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	H13p1
Valor mínimo	0
Primer cuantil	0.2217
Mediana	0.2998
Media	0.3050
Tercer cuantil	0.3750
Valor máximo	1
Desviación estándar	0.1339978
Coeficiente de asimetría	0.8013251
Coeficiente de Kurtosis	5.3032149

Cuadro 14: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *H13p1*

Los estadísticos nos dan nuevamente una idea de una distribución con un centro denso desplazado a la izquierda de la distribución, la cual presenta también una amplia dispersión de los datos. Esto indicaría nuevamente la presencia de *outliers* especialmente situados a la derecha de la distribución.

Se verifica esta información de forma gráfica con ayuda de un diagrama *boxplot* y un histograma:

```

1 # H13p1
2
3 # Diagrama boxplot
4 graf <- ggplot(house, aes(x='H13p1', y=H13p1)) +
5   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
6     outlier.shape=1) +
7   labs(caption='center', y=' ', x=' ')
8 graf
9
10 # Histograma
11 graf <- ggplot(house, aes(H13p1)) +
12   geom_histogram(binwidth=0.0005, color='gray58') +
13   labs(y='frecuencia absoluta')
14 graf

```

Script 8: Conjunto de sentencias para dibujar y visualizar un diagrama *boxplot* y un histograma de la variable *H13p1*

Se muestra el diagrama *boxplot*:

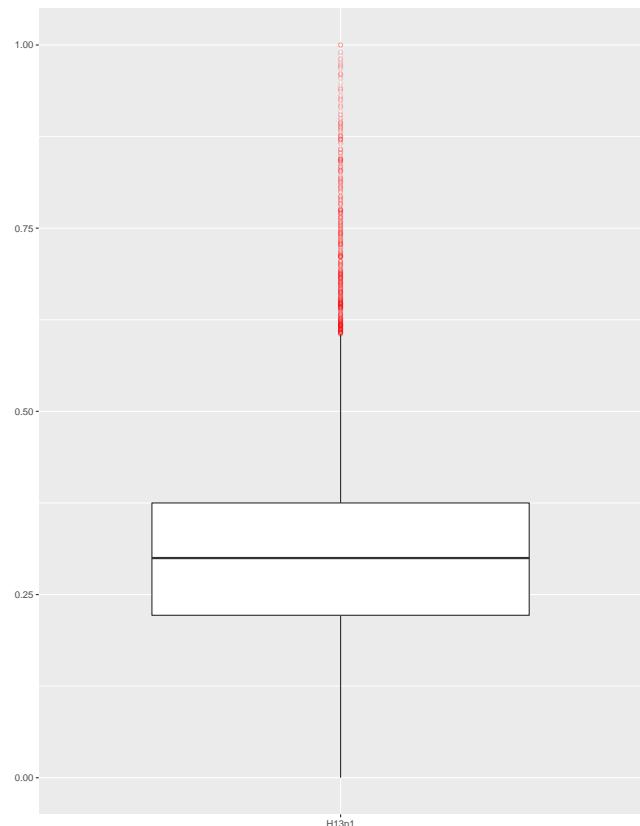


Figura 30: Histograma de la variable $H13p1$

Y el histograma con la forma de la distribución:

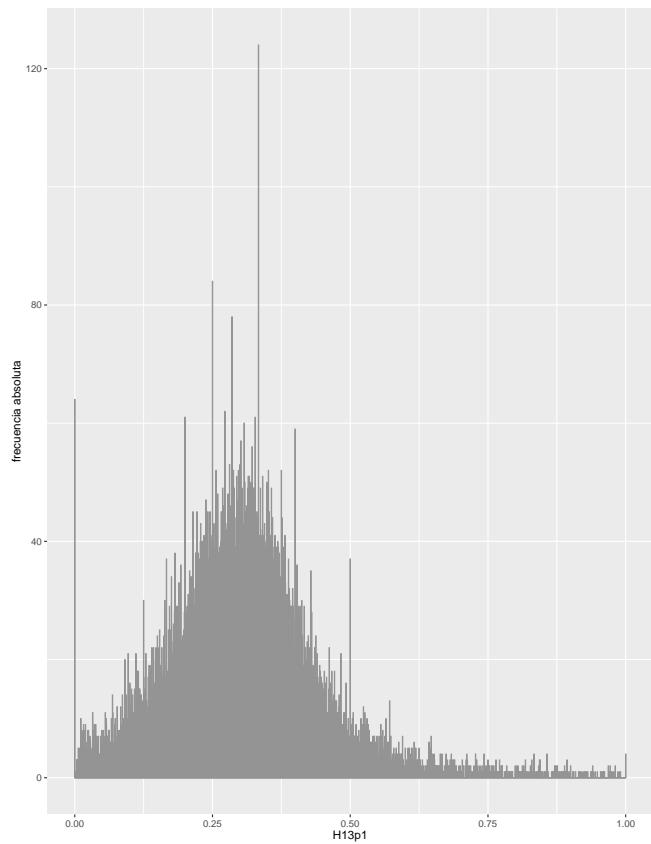


Figura 31: Histograma de la variable $H13p1$

En el histograma podemos observar como, además del centro de la distribución, la densidad de la distribución es elevada para el valor 0.

- **H18pA:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	H18pA
Valor mínimo	0
Primer cuantil	0.05277
Mediana	0.08696
Media	0.10738
Tercer cuantil	0.13793
Valor máximo	1
Desviación estándar	0.09006532
Coeficiente de asimetría	2.613881
Coeficiente de Kurtosis	17.009835

Cuadro 15: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *H18pA*

Los estadísticos nos vuelven a dar una idea de una distribución con un centro denso y muy estrecho desplazado a la izquierda de la distribución. La distribución presenta gran dispersión de sus datos respecto de este centro de distribución, lo que podría llevar a la aparición de *outliers*, localizados especialmente a la derecha de la distribución.

Se verifica esta información de forma gráfica con ayuda de un diagrama *boxplot* y un histograma:

Se muestran a continuación ambos diagramas:

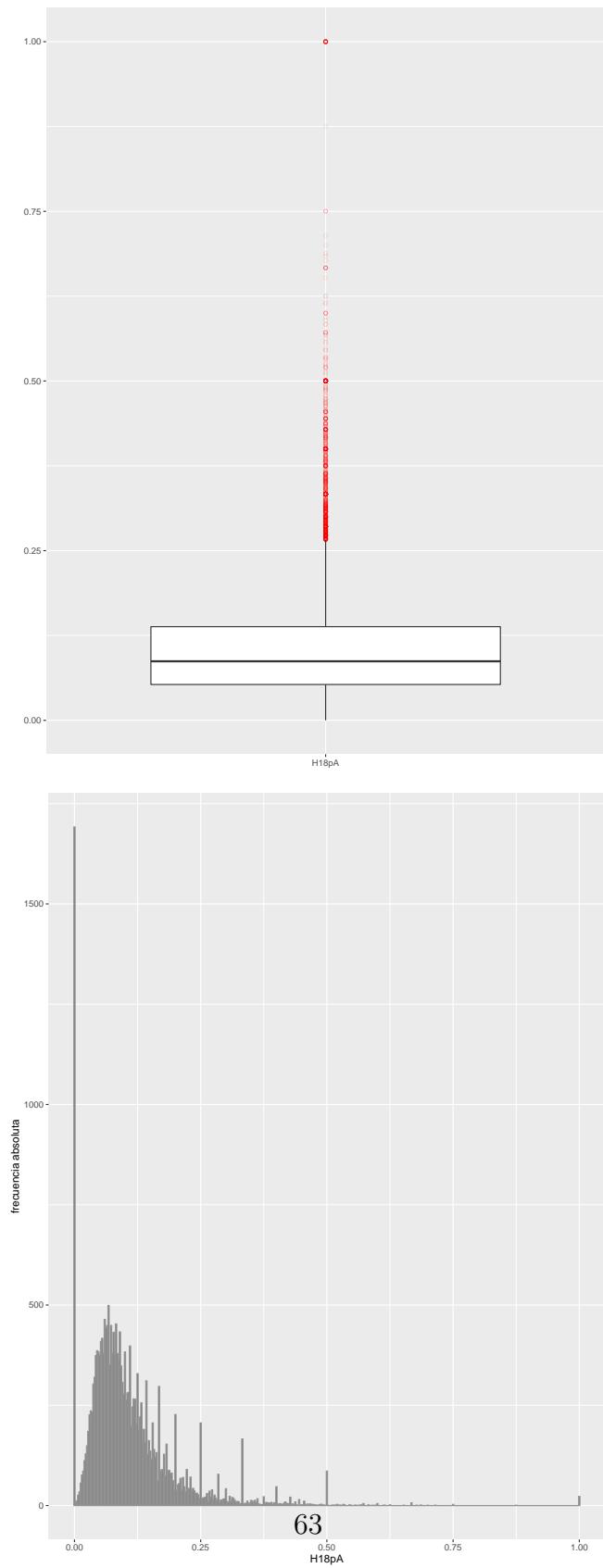


Figura 32: Representación del diagrama *boxplot* (diagrama superior) e histograma (diagrama inferior) de la variable $H18pA$

Observando la forma de la distribución reflejada en el histograma, se vuelve a apreciar que, además del centro de la distribución, organizado en torno a la mediana de la distribución, la distribución presenta una zona de mayor densidad para el valor 0.

- **H40p4:** Los estadísticos calculados para este atributo se recogen en la siguiente tabla:

	H40p4
Valor mínimo	0
Primer cuantil	0.2432
Mediana	0.5
Media	0.4916
Tercer cuantil	0.75
Valor máximo	1
Desviación estándar	0.3316551
Coeficiente de asimetría	-0.02660079
Coeficiente de Kurtosis	1.90407684

Cuadro 16: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *H40p4*

Los estadísticos nos dan una idea de que la distribución se halla centrada en el centro de la distribución (algo desplazada a la derecha) y que presenta cierta dispersión de los datos respecto de este centro.

Para analizar más en detalle este fenómeno, se estudia el diagrama *boxplot* y el histograma de esta distribución:

Se verifica esta información de forma gráfica con ayuda de un diagrama *boxplot* y un histograma:

Se muestran a continuación ambos diagramas:

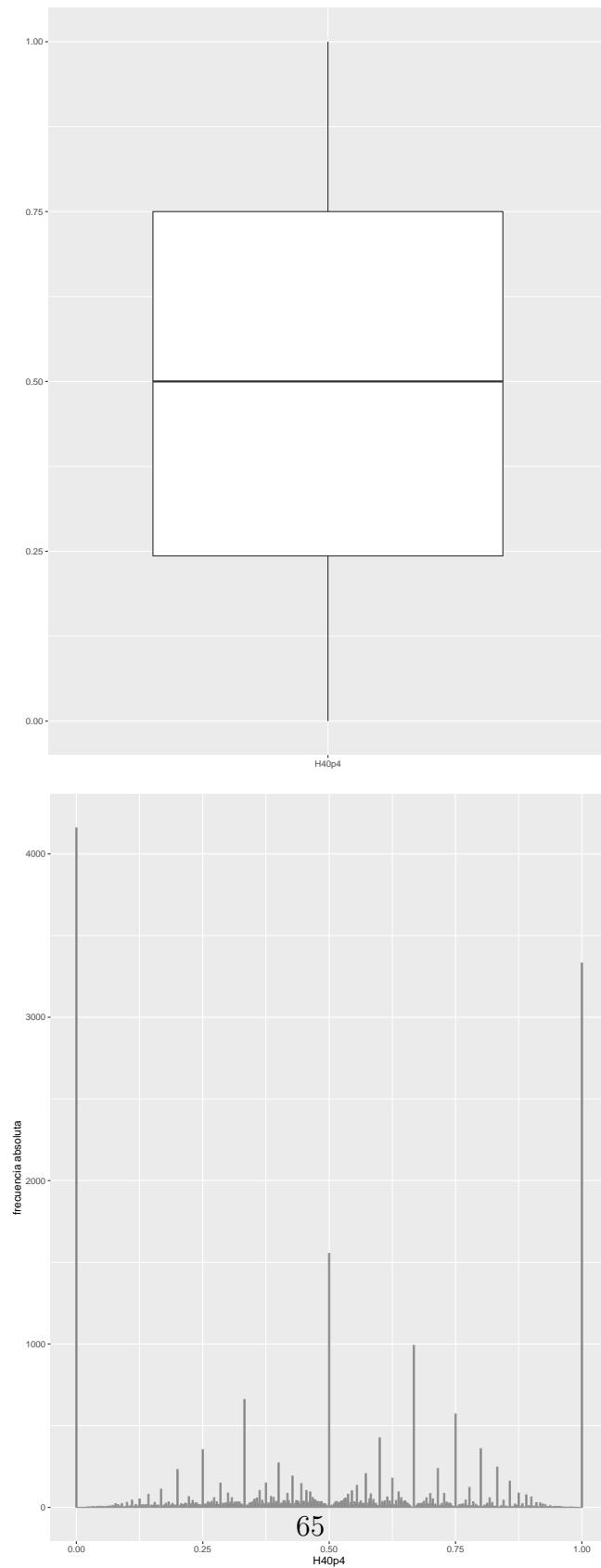


Figura 33: Representación del diagrama *boxplot* (diagrama superior) e histograma (diagrama inferior) de la variable $H40p4$

A partir de la forma de la distribución, observamos detalles muy importantes acerca de la misma: las regiones de mayor densidad se encuentran principalmente para los valores 0, 1 y en la mediana de la distribución. Por otra parte, la distribución presenta otras regiones menos densas a lo largo del dominio pero centradas en la mediana, razón por la cual se obtienen los estadísticos calculados.

- **Price:** Para la variable dependiente, los estadísticos calculados se recogen en la siguiente tabla:

	Price
Valor mínimo	0
Primer cuantil	21000
Mediana	33200
Media	50074
Tercer cuantil	56100
Valor máximo	500001
Desviación estándar	52843.48
Coeficiente de asimetría	3.754873
Coeficiente de Kurtosis	23.322199

Cuadro 17: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *Price*

Los estadísticos nos dan una idea de que el centro de la distribución se halla profundamente desplazado a la izquierda y existe dispersión de los datos respecto de este centro.

Se analiza gráficamente la distribución de esta variable mediante un diagrama *boxplot* y un histograma:

Se muestran a continuación ambos diagramas:

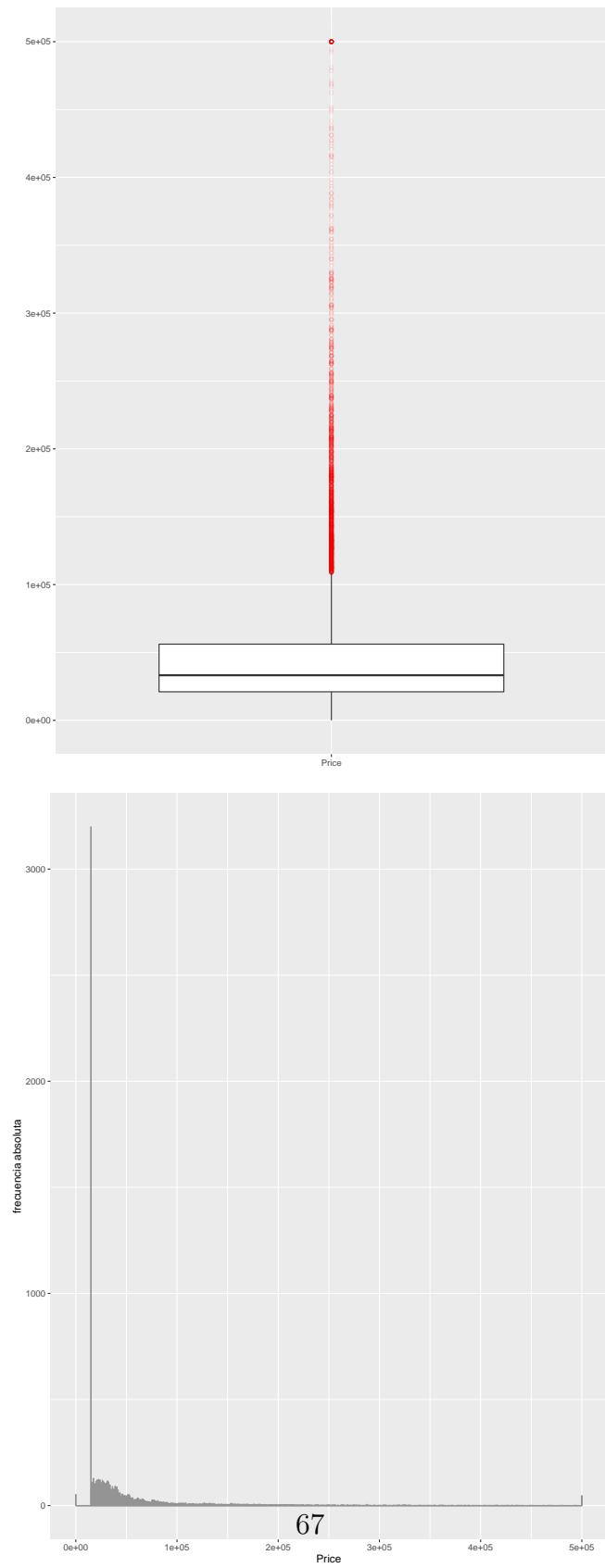


Figura 34: Representación del diagrama *boxplot* (diagrama superior) e histograma (diagrama inferior) de la variable *Price*

Analizando el histograma de la distribución, se observa que sufre un incremento brusco de su densidad, alcanzando su valor máximo en la distribución al aproximarse a la mediana de la distribución, tras el cual, la densidad sufre un decremento brusco tras el cual, comienza a decremetarse de una forma más suave.

Interpretando estos resultados, podemos afirmar que la mayor parte de las regiones recogidas en el dataset, tienen un precio medio de, aproximadamente 33200 unidades monetarias, por debajo de este valor, resulta infrecuente encontrar regiones con ese precio. Por su parte, si consideramos precios medios superiores, el número de regiones es bastante inferior y su recuencia decrece progresivamente. Por último, para valores superiores a 56100 unidades monetarias podríamos considerar que sólo se encuentran algunas regiones muy concretas con precios extremadamente altos.

En otras palabras, el precio de las viviendas de *dataset* oscilaría entre las 33200 unidades monetarias (más frecuentemente) y las 56100 unidades monetarias (menos frecuente) dándose ciertas regiones en las que el precio es extremadamente elevado y menos regiones en las que el precio sea extremadamente bajo.

2.2.2. Estudio de las relaciones entre variables

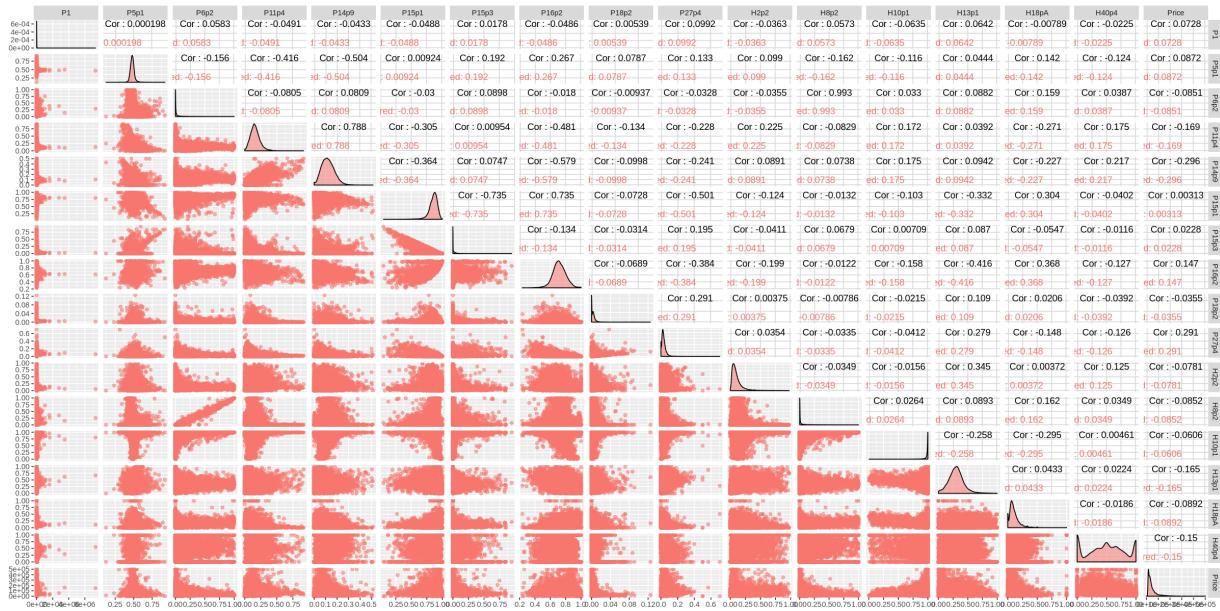
Una vez se conocen todas las variables, así como las distribuciones de sus datos, se decide estudiar las relaciones existentes entre las mismas con el objetivo de obtener más información sobre el conjunto de datos.

En este sentido, el análisis se orientará con la finalidad de determinar si existe alguna variable que pueda ser explicada por otra mediante una relación lineal o, una relación no lineal conocida y, sobre todo, para determinar la existencias de dependencias entre las variables que deban de ser tenida en la elaboración de modelos.

Primeramente, se analizará de forma gráfica las relaciones existentes entre los pares de las variables existentes. De este modo, para cada par de variables, se elabora un diagrama *scatterplot* que permitirá, de forma visual, determinar si existe alguna relación conocida entre las dos variables y encontrar detalles significativos.

En la siguiente gráfica, se recoge organizada en una cuadícula, los dia-

gramas *scatterplot* de cada par de variables. Conviene destacar que, en la diagonal superior de la cuadrícula, en lugar de un diagrama *scatterplot*, se muestra el **coeficiente de correlación lineal de Kendall** calculado para cada par de variables.



Se decide también cuantificar de forma analítica las relaciones lineales entre las variables, para lo cual, y puesto que muchas de las variables del *dataset* no tienden a la normalidad, se usará como métrica el coeficiente de correlación de *Kendall*, el cual se muestra en la siguiente figura:

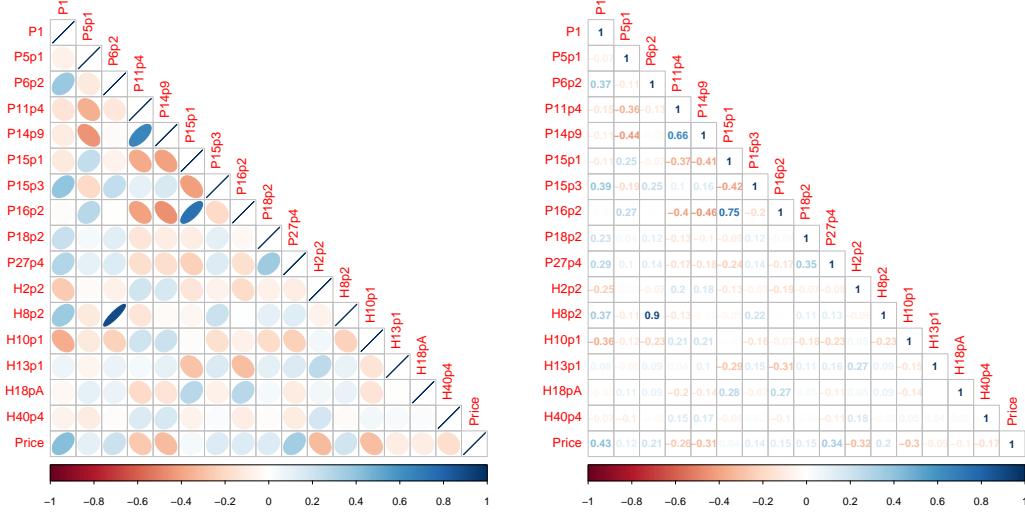


Figura 35: A la derecha se muestra una matriz triangular inferior donde cada celda muestra el valor del coeficiente de Kendall asociado a cada par de variables, nótese que un color cercano a azul oscuro indica una correlación lineal directa, mientras que un color rojo indica una correlación lineal inversa. A la izquierda se representa otro correlograma donde el valor del coeficiente ha sido sustituido por una esfera cuyo color y forma muestran el grado de correlación lineal entre las variables.

Con ayuda del anterior gráfico, analizamos el siguiente caso:

Al incrementar el tamaño de la población, los porcentajes demográficos medidos por el resto de variables sobre la población tienden a un valor central, mientras que ante un tamaño de población total baja, estos porcentajes oscilan entre sus valores máximos y mínimos:

Si observamos la relación entre las variables $P1$ y $P5p1$, que se representa más detalladamente en la figura 36, observamos que para un tamaño de población próximo a 0, el porcentaje de población masculina oscila entre sus valores máximos y mínimos. Al aumentar el tamaño de la población, el porcentaje de la población masculina reduce su oscilación y tiende a 0.5.

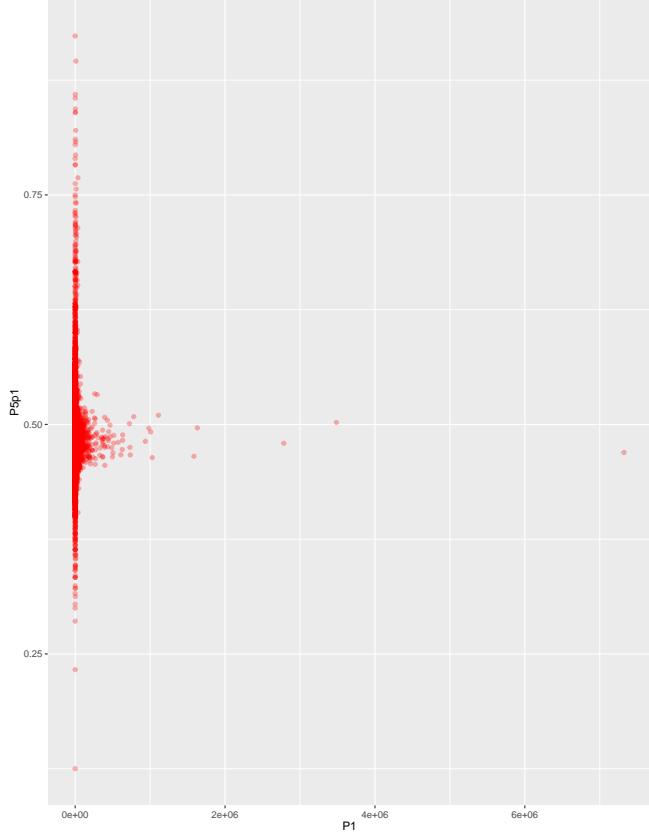


Figura 36: Diagrama de puntos que relaciona la variable P5p1 frente a P1

En otras palabras, al aumentar el tamaño de la población, se acentúa la tendencia demográfica que establece un porcentaje similar de hombres y mujeres que conforman la población.

Para el resto de variables, se observa un fenómeno similar: al aumentar el tamaño de la población, los porcentajes disminuyen su oscilación tiendiendo a un valor, lo cual indicaría que en la población existen tendencias demográficas a los que la población converge con un tamaño suficientemente grande de población.

De este modo, la relación entre la variable $P11p4$ y $P1$ nos indica que el porcentaje de la población con 5 años tiende al 15 % (aproximadamente la mediana de la distribución de $P11p4$); la relación entre la variable $P14p9$ y $P1$ nos indica que el porcentaje de mujeres viudas tiende al 10 %; la relación entre la variable $P15p1$ y $P1$ indica que el porcentaje de la población que vive en hogares con un cabeza de familia tiende al 80 %; la relación entre

la variable $P27p4$ y $P1$ indicaría que la tendencia en el porcentaje de los hogares con 5 miembros es próxima a 0; la relación entre la variable $H2p2$ y $P1$ indicaría que la tendencia en el porcentaje de número de unidades inmobiliarias vacías es próxima al 10 %; y la relación entre la variable $H18pA$ y $P1$ indicaría que la tendencia en el porcentaje de personas poseedoras de su unidad inmobiliaria es del 10 %, lo que indicaría que en la mayoría de las personas que residen en una unidad inmobiliaria no son poseedoras de la misma.

Para el resto de variables, la tendencia es menos clara y se observa una mayor oscilación.

2.3. Elaboración de modelos predictivos

Una vez llevado conocidos todos los detalles más significativos de la información contenida en este *dataset*, se tratará la elaboración de modelos predictivos a partir de la información del mismo.

Para la elaboración de modelos que permitan explicar la variable dependiente *Price*, en primer lugar, se comenzará estudiando y elaborando modelos lineales simples, a partir de los cuales, se elaborarán otros modelos progresivamente más complejos, modelos no lineales y también, modelos basados en *k-NN*.

2.3.1. Elaboración de modelos lineales simples

En primer lugar, se tratará de estudiar y generar modelos lineales simples para explicar la variable dependiente en función de las demás.

El estudio de estos modelos lineales, se enfoca principalmente, para la detección de variables independientes que, por si mismas, expliquen un buen porcentaje de la variable dependiente, lo cual resultará significativo para elaborar otros modelos lineales compuestos y modelos no lineales.

Se desarrollan los siguientes modelos lineales:

```

1 # Modelos lineales
2 lineal.simple.P1 <- lm(Price ~ P1, data=house)
3 lineal.simple.P5p1 <- lm(Price ~ P5p1, data=house)
4 lineal.simple.P6p2 <- lm(Price ~ P6p2, data=house)
5 lineal.simple.P11p4 <- lm(Price ~ P11p4, data=house)
6 lineal.simple.P14p9 <- lm(Price ~ P14p9, data=house)
7 lineal.simple.P14p9 <- lm(Price ~ P14p9, data=house)
```

```

8 lineal.simple.P15p1 <- lm(Price ~ P15p1, data=house)
9 lineal.simple.P15p3 <- lm(Price ~ P15p3, data=house)
10 lineal.simple.P16p2 <- lm(Price ~ P16p2, data=house)
11 lineal.simple.P18p2 <- lm(Price ~ P18p2, data=house)
12 lineal.simple.P27p4 <- lm(Price ~ P27p4, data=house)
13 lineal.simple.H2p2 <- lm(Price ~ H2p2, data=house)
14 lineal.simple.H8p2 <- lm(Price ~ H8p2, data=house)
15 lineal.simple.H10p1 <- lm(Price ~ H10p1, data=house)
16 lineal.simple.H13p1 <- lm(Price ~ H13p1, data=house)
17 lineal.simple.H18pA <- lm(Price ~ H18pA, data=house)
18 lineal.simple.H40p4 <- lm(Price ~ H40p4, data=house)
19
20 # Obtener información sobre los modelos
21 summary(lineal.simple.P1)
22 evaluate.rmse.house(lineal.simple.P1)
23
24 summary(lineal.simple.P5p1)
25 evaluate.rmse.house(lineal.simple.P5p1)
26
27 summary(lineal.simple.P6p2)
28 evaluate.rmse.house(lineal.simple.P6p2)
29
30 summary(lineal.simple.P11p4)
31 evaluate.rmse.house(lineal.simple.P11p4)
32
33 summary(lineal.simple.P14p9)
34 evaluate.rmse.house(lineal.simple.P14p9)
35
36 summary(lineal.simple.P15p1)
37 evaluate.rmse.house(lineal.simple.P15p1)
38
39 summary(lineal.simple.P15p3)
40 evaluate.rmse.house(lineal.simple.P15p3)
41
42 summary(lineal.simple.P16p2)
43 evaluate.rmse.house(lineal.simple.P16p2)
44
45 summary(lineal.simple.P18p2)
46 evaluate.rmse.house(lineal.simple.P18p2)
47
48 summary(lineal.simple.P27p4)
49 evaluate.rmse.house(lineal.simple.P27p4)
50
51 summary(lineal.simple.H2p2)
52 evaluate.rmse.house(lineal.simple.H2p2)
53
54 summary(lineal.simple.H8p2)
55 evaluate.rmse.house(lineal.simple.H8p2)
56

```

```

57 summary(lineal.simple.H10p1)
58 evaluate.rmse.house(lineal.simple.H10p1)
59
60 summary(lineal.simple.H13p1)
61 evaluate.rmse.house(lineal.simple.H13p1)
62
63 summary(lineal.simple.H18pA)
64 evaluate.rmse.house(lineal.simple.H18pA)
65
66 summary(lineal.simple.H40p4)
67 evaluate.rmse.house(lineal.simple.H40p4)

```

Script 9: Conjunto de sentencias para la elaboración de modelos lineales y obtención de información de los mismos, además del cómputo del error RMSE para cada uno de ellos

Para cada modelo lineal elaborado, se evalúan en la siguiente table:

Modelo	R^2	R^2 Ajustado	RMSE	p-value
$Price \sim P1$	0.005298	0.005254	52702.16	2.2e-16
$Price \sim P5p1$	0.007606	0.007563	52640.96	2.2e-16
$Price \sim P6p2$	0.007248	0.007205	52650.46	2.2e-16
$Price \sim P11p4$	0.02855	0.02851	52082.59	2.2e-16
$Price \sim P14p9$	0.08733	0.08715	50482.2	2.2e-16
$Price \sim P15p1$	9.786e-06	-3.413e-05	52842.06	0.6371
$Price \sim P15p3$	0.0005219	0.0005632	52828.52	2.2e-16
$Price \sim P16p2$	0.02175	0.02171	52264.44	2.2e-16
$Price \sim P18p2$	0.001217	0.001213	52809.1	8.667e-08
$Price \sim P27p4$	0.08452	0.08448	50559.88	2.2e-16
$Price \sim H2p2$	0.006099	0.006055	52680.94	2.2e-16
$Price \sim H8p2$	0.007261	0.007217	52650.13	2.2e-16
$Price \sim H10p1$	0.003676	0.003632	52745.12	2.2e-16
$Price \sim H13p1$	0.02719	0.02715	52119.01	2.2e-16
$Price \sim H18pA$	0.007961	0.007917	52631.56	2.2e-16
$Price \sim H40p4$	0.02239	0.02235	52247.27	2.2e-16

Cuadro 18: Listado de los modelos lineales simples elaborados para cada uno de los cuales se evalúa: coeficiente de determinación R^2 , R^2 Ajustado, el error RMSE y el p-value asociado al test de Wall

Los modelos obtenidos se representan de forma gráfica en los siguientes diagramas *scatterplot* en los que, para cada variable independiente, se ha representado la nube de puntos resultante de comparar la variable *Price* con ella misma junto con el correspondiente modelo lineal elaborado.

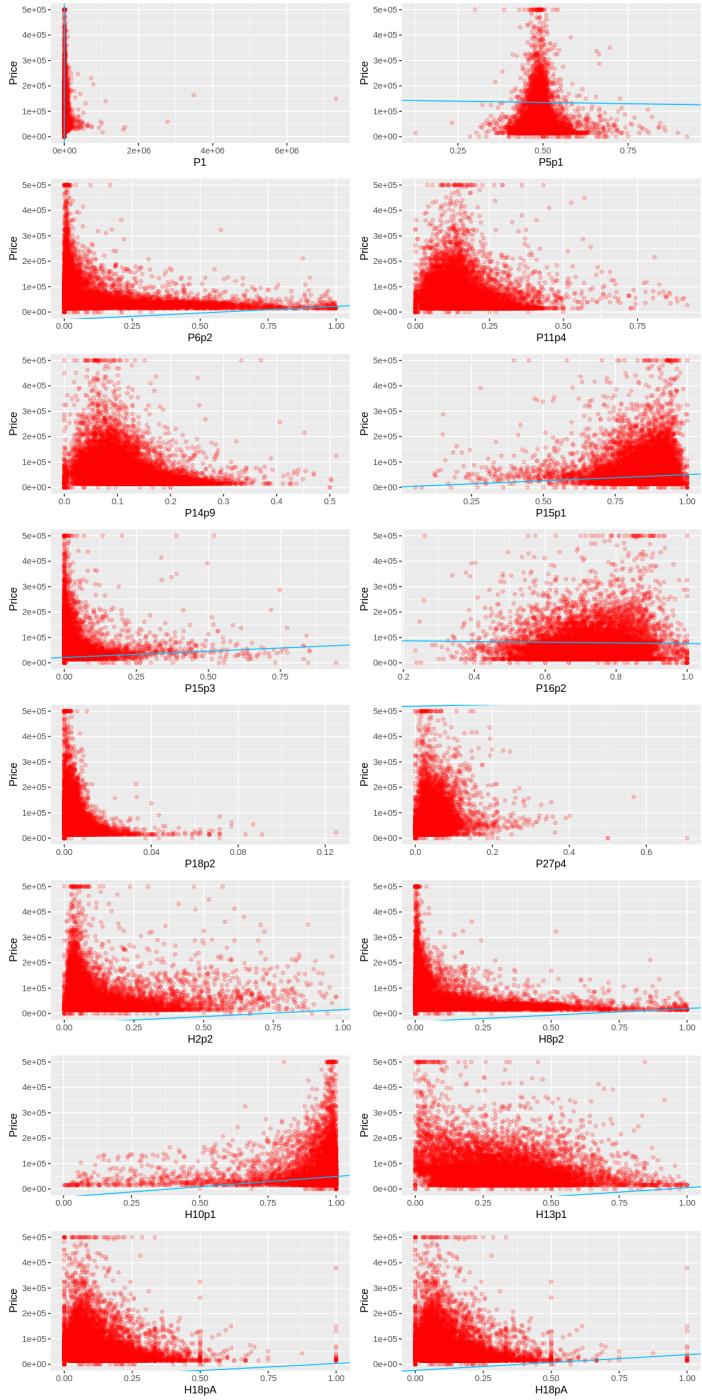


Figura 37: Representaciones de la variable *Price* frente a cada una de las variables independientes del *dataset*. Para cada una de las representaciones, se representa la nube de puntos en color rojo y la línea que representa el modelo lineal generado en cada caso en color azul claro

Evaluando las métricas de evaluación de los modelos, observamos que, como se esperaba, ningún modelo constituye una buena aproximación para predecir la variable dependiente, pues todos los modelos presentan valores de R^2 y R^2 Ajustados próximos a 0, con un error *RMSE* (*Root Mean Squared Error*).

Por su parte, para el modelo que explica *Price* en función de *P15pq*, observamos que el valor *p-value* del test de Wall realizado para determinar si el coeficiente de la variable independiente es igual a 0 nos da un valor significativo que nos llevaría a aceptar la hipótesis nula de este test, y afirmar que en este modelo, la variable independiente puede ser desestimada.

Finalmente, se seleccionan los 5 regresores lineales que ofrecen un mejor rendimiento en función de las métricas evaluadas y recogidas en la tabla 18:

Ranking	Modelo	R^2	R^2 Ajustado	RMSE	p-value
1º	$Price \sim P27p4$	0.08452	0.08448	50559.88	$< 2,2e - 16$
2º	$Price \sim P11p4$	0.02855	0.02851	52082.59	$< 2,2e - 16$
3º	$Price \sim H13p1$	0.02719	0.02715	52119.01	$< 2,2e - 16$
4º	$Price \sim H40p4$	0.02239	0.02235	52247.27	$< 2,2e - 16$
5º	$Price \sim P16p2$	0.02175	0.02171	52264.44	$< 2,2e - 16$

Cuadro 19: Ranking de los 5 modelos lineales que ofrecen mejores resultados dadas las métricas evaluadas. Para cada modelo se vuelve a mostrar las métricas: coeficiente de determinación R^2 , R^2 Ajustado, el error *RMSE* y el *p-value* asociado al test de Wall

Por último, comparamos estos 5 modelos evaluando además la capacidad de generalización de los mismos, para lo cual, se va a medir el error *RMSE* cometido por cada uno mediante validación cruzada con *10-fold*:

Ranking	Modelo	R^2	R^2 Ajustado	RMSE	5-fold RMSE
1º	$Price \sim P27p4$	0.08452	0.08448	50559.88	50529.49
2º	$Price \sim P11p4$	0.02855	0.02851	52082.59	52049.82
3º	$Price \sim H13p1$	0.02719	0.02715	52119.01	52094.35
4º	$Price \sim H40p4$	0.02239	0.02235	52247.27	52219.24
5º	$Price \sim P16p2$	0.02175	0.02171	52264.44	52237.49

Cuadro 20: Ranking de los 5 modelos lineales que ofrecen mejores resultados dadas las métricas evaluadas. Para cada modelo se vuelve a mostrar las métricas: coeficiente de determinación R^2 , R^2 Ajustado, el error $RMSE$ y el error $RMSE$ medido mediante validación cruzada

Las métricas evaluadas no presentan discrepancias en la comparación de modelos. Se observa que **el modelo lineal $Price = b_0 + b_1 * P27p4$ es el que ofrece mejores resultados** para todas las métricas evaluadas.

2.3.2. Elaboración de modelos lineales compuestos

Una vez realizado un primer estudio de las variables que más determinan el valor de $Price$, se lleva a cabo el estudio y elaboración de modelos lineales y no lineales que permitan explicar de forma más óptima la variable dependiente $Price$.

En primer lugar, se construirá un modelo lineal compuesto (*lineal.model.fit1*) considerando todas las variables dependientes del *dataset* y se evaluará su rendimiento con las métricas empleadas en el apartado anterior.

```

1 # Modelo lineal compuesto por todas las variables del
  dataset
2 lineal.model <- lm(Price ~ ., data=house)
3
4 summary(lineal.model)
5 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
  model), fill=T)
6 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
  : ', mean(sapply(1:5, run_k_fold,
7
  , nombre
  , lineal.model), fill=T))

```

Script 10: Conjunto de sentencias para la elaboración y evaluación de un modelo lineal compuesto por todas las variables dependientes

El resultado devuelve la siguiente salida:

Call:

```

lm(formula = Price ~ ., data = house)

Residuals:
    Min      1Q  Median      3Q     Max 
-483860 -19545   -7505    5775  450186 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.403e+05  1.709e+04 19.912 < 2e-16 ***
P1          3.760e-02  4.625e-03  8.130 4.51e-16 ***
P5p1        -1.126e+05 1.192e+04 -9.439 < 2e-16 *** 
P6p2         7.721e+04  1.790e+04  4.314 1.61e-05 ***
P11p4        6.031e+04  7.088e+03  8.509 < 2e-16 *** 
P14p9        -1.805e+05 1.126e+04 -16.027 < 2e-16 *** 
P15p1        -5.012e+05 2.547e+04 -19.676 < 2e-16 *** 
P15p3        -4.360e+05 2.275e+04 -19.162 < 2e-16 *** 
P16p2         3.858e+05  1.433e+04  26.931 < 2e-16 *** 
P18p2        -1.349e+06 6.269e+04 -21.513 < 2e-16 *** 
P27p4         5.205e+05  1.902e+04  27.364 < 2e-16 *** 
H2p2          -1.132e+03 3.283e+03 -0.345     0.73  
H8p2          -8.338e+04 1.927e+04 -4.326 1.53e-05 *** 
H10p1         -4.712e+04 3.379e+03 -13.945 < 2e-16 *** 
H13p1         -7.006e+04 2.901e+03 -24.153 < 2e-16 *** 
H18pA         -6.027e+04 3.900e+03 -15.455 < 2e-16 *** 
H40p4         -8.330e+03 9.458e+02 -8.807 < 2e-16 *** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45440 on 22767 degrees of freedom
Multiple R-squared:  0.2611, Adjusted R-squared:  0.2606 
F-statistic: 502.9 on 16 and 22767 DF,  p-value: < 2.2e-16

RMSE del modelo: 45422.33
Error RMSE del modelo lineal compuesto sobre 5-fold: 45504.1

```

Se aprecia que el modelo originado ofrece un rendimiento superior a cualquiera de los modelos lineales simples elaborados en el sub-epígrafe anterior. No obstante, se sigue observando que el rendimiento no consigue explicar un porcentaje deseable de la variable dependiente (valores de coeficiente de determinación y coeficiente de determinación no altos), además de un error RMSE alto.

Por otro lado, el *p-value* del test de Wall realizado para el coeficiente calculado para la variable textitH2p2 presenta un valor muy significativo, lo que lleva a pensar que resulta redundante y que podría ser eliminado:

```

1 # Modelo lineal sin H2p2
2 lineal.model.fit1 <- lm(Price ~ . - H2p2, data=house)
3
4 summary(lineal.model.fit1)
5 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
     model.fit1), fill=T)
6 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
     : ', mean(sapply(1:5, run_k_fold,
7
     nombre, lineal.model.fit1), fill=T))

```

Script 11: Conjunto de sentencias para la elaboración y evaluación de un modelo lineal compuesto por todas las variables dependientes a excepción de *H2p2*

Se obtiene la siguiente salida:

Call:

```
lm(formula = Price ~ . - H2p2, data = house)
```

Residuals:

Min	1Q	Median	3Q	Max
-483922	-19538	-7476	5773	450224

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.403e+05	1.709e+04	19.910	< 2e-16 ***
P1	3.768e-02	4.621e-03	8.154	3.70e-16 ***
P5p1	-1.132e+05	1.176e+04	-9.628	< 2e-16 ***
P6p2	7.717e+04	1.790e+04	4.312	1.63e-05 ***
P11p4	5.974e+04	6.889e+03	8.672	< 2e-16 ***
P14p9	-1.800e+05	1.117e+04	-16.114	< 2e-16 ***
P15p1	-5.004e+05	2.536e+04	-19.730	< 2e-16 ***
P15p3	-4.351e+05	2.260e+04	-19.250	< 2e-16 ***
P16p2	3.855e+05	1.429e+04	26.969	< 2e-16 ***
P18p2	-1.348e+06	6.268e+04	-21.511	< 2e-16 ***
P27p4	5.210e+05	1.898e+04	27.451	< 2e-16 ***
H8p2	-8.335e+04	1.927e+04	-4.324	1.54e-05 ***
H10p1	-4.720e+04	3.371e+03	-14.002	< 2e-16 ***
H13p1	-7.037e+04	2.757e+03	-25.527	< 2e-16 ***

```

H18pA      -6.036e+04  3.891e+03 -15.514  < 2e-16 ***
H40p4      -8.370e+03  9.385e+02 -8.918  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45440 on 22768 degrees of freedom
Multiple R-squared:  0.2611, Adjusted R-squared:  0.2606
F-statistic: 536.4 on 15 and 22768 DF,  p-value: < 2.2e-16

RMSE del modelo: 45422.45
Error RMSE del modelo lineal compuesto sobre 5-fold: 45498.44

```

Se observa que el rendimiento de este modelo es similar al rendimiento del modelo original y que, por tanto, la variable $H2p2$ no aportaba información.

A partir de este modelo, se concebirán otros modelos incluyendo términos no lineales con la finalidad de mejorar el rendimiento del mismo.

Como primera modificación, probamos a añadir un término con la variable $P27p4$ elevada a diferentes potencias:

```

1 # Modelo no lineal con P27p4 al cuadrado
2 lineal.model.fit2 <- lm(Price~.-H2p2+I(P27p4^2), data=
  house)
3
4 summary(lineal.model.fit2)
5 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
  model.fit2), fill=T)
6 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
  : ', mean(sapply(1:5, run_k_fold,
7
  nombre, lineal.model.fit2), fill=T))

```

Script 12: Conjunto de sentencias para la elaboración de un modelo lineal compuesto con un término que incluye la variable $P17p4$ elevada al cuadrado

Considerando el siguiente modelo $lineal.model.fit2$:

$$Price = b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	p-value	RMSE 5-fold
0.2894	0.2899	44530.35	< 2,2e - 16	44629.42

Cuadro 21: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P27p4$ elevado al cuadrado

Consideramos ahora el siguiente modelo *lineal.model.fit3*:

$$\begin{aligned} Price = & b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + \\ & b_6 * P15p1 + b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\ & b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^3 \end{aligned}$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.2794	0.2789	44856.75	45083.58

Cuadro 22: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P27p4$ elevado a 3

Observamos que el rendimiento de este modelo es ligeramente inferior al anterior.

Por último, dado el modelo *lineal.model.fit4*:

$$\begin{aligned} Price = & b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\ & b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\ & b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^4 \end{aligned}$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.2725	0.272	45071.03	45432.03

Cuadro 23: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P27p4$ elevado a 4

El rendimiento del modelo vuelve a ser inferior al modelo que consideraba el término con $P27p4$ elevado al cuadrado, por lo tanto, se decide tomar modelo con el término que incluye $P27p4$ al cuadrado.

Seguidamente, se deciden introducir otros términos con variables al cuadrado. Se inserta un término con $P11p4$ al cuadrado, dando lugar al modelo *lineal.model.fit5*:

$$Price = b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * P11p4^2$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.29	0.2894	44526.54	44628.2

Cuadro 24: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P11p4$ elevado a 2

Añadimos un término con $H13p1$ al cuadrado, dando lugar al modelo *lineal.model.fit6*:

$$Price = b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * P11p4^2 + b_{18} * H13p1^2$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.3201	0.3196	43570.5	43719.26

Cuadro 25: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H13p1$ elevado a 2 y eliminando el término con $P11p4$ al cuadrado

Por su parte, para este último modelo, se obtuvo la siguiente salida:

Call:

```
lm(formula = Price ~ . - H2p2 + I(P27p4^2) + I(P11p4^2) + I(H13p1^2),
   data = house)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-211862	-19267	-5374	7992	460257

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.403e+05	1.643e+04	20.713	< 2e-16 ***
P1	2.977e-02	4.438e-03	6.708	2.02e-11 ***
P5p1	-1.383e+05	1.133e+04	-12.205	< 2e-16 ***
P6p2	7.536e+04	1.718e+04	4.387	1.16e-05 ***
P11p4	7.476e+04	1.210e+04	6.179	6.55e-10 ***
P14p9	-1.356e+05	1.089e+04	-12.456	< 2e-16 ***
P15p1	-4.255e+05	2.520e+04	-16.887	< 2e-16 ***
P15p3	-3.513e+05	2.233e+04	-15.737	< 2e-16 ***
P16p2	3.367e+05	1.453e+04	23.175	< 2e-16 ***
P18p2	-1.561e+06	6.112e+04	-25.536	< 2e-16 ***
P27p4	1.002e+06	2.297e+04	43.610	< 2e-16 ***
H8p2	-7.734e+04	1.850e+04	-4.180	2.92e-05 ***
H10p1	-5.126e+04	3.241e+03	-15.816	< 2e-16 ***
H13p1	-2.912e+05	7.313e+03	-39.816	< 2e-16 ***
H18pA	-5.312e+04	3.773e+03	-14.080	< 2e-16 ***
H40p4	-4.939e+03	9.057e+02	-5.453	5.01e-08 ***
I(P27p4^2)	-2.233e+06	6.807e+04	-32.804	< 2e-16 ***
I(P11p4^2)	-2.953e+04	1.900e+04	-1.554	0.12
I(H13p1^2)	2.922e+05	9.195e+03	31.780	< 2e-16 ***

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

Residual standard error: 43590 on 22765 degrees of freedom

Multiple R-squared: 0.3201, Adjusted R-squared: 0.3196

F-statistic: 595.5 on 18 and 22765 DF, p-value: < 2.2e-16

RMSE del modelo: 43570.5

Error RMSE del modelo lineal compuesto sobre 5-fold: 43719.26

Para el término que incluía la variable $P11p4$ al cuadrado, el test de Wall realizado indica que, muy significativamente, el coeficiente de este término sea igual a 0, por lo tanto, puede ser eliminado.

Se procede a eliminar otro modelo (*lineal.model.fit7*) eliminando este término:

$$Price = b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.3201	0.3196	43572.81	43718.63

Cuadro 26: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H13p1$ elevado a 2

El rendimiento de este modelo es similar al modelo *lineal.model.fit6*.

Se prueba ahora a añadir un término con $H40p4$ elevado al cuadrado, dando lugar al modelo *lineal.model.fit8*:

$$Price = b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.3375	0.337	43010.9	43126.04

Cuadro 27: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H40p4$ elevado a 2

El rendimiento del modelo ha mejorado con respecto al anterior, se decide continuar añadiendo términos con las variables del *dataset* elevados al cuadrado. Se añade ahora otro término con $P16p2$ al cuadrado dando lugar al modelo *lineal.model.fit9*:

$$Price = b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + b_{19} * P16p2^2$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.3393	0.3388	42950.6	43072.89

Cuadro 28: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P16p2$ elevado a 2

Nuevamente, el rendimiento del modelo ha sufrido una pequeña mejora. Hasta ahora, sólo se han incluído las variables que se consideraban más influyentes dados los modelos lineales simples elaborados en el epígrafe anterior, se decide repetir esta operación con el resto de variables del *dataset*: Se añade ahora otro término con la variable $P1$ al cuadrado en el modelo *lineal.model.fit10*:

$$\begin{aligned} Price = & b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\ & b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\ & b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + \\ & b_{19} * P16p2^2 + b_{20} * P1^2 \end{aligned}$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.34	0.3394	42928.75	43496.06

Cuadro 29: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P1$ elevado a 2

El rendimiento del modelo vuelve a mejorar. Se repite este proceso con la variable $P5p1$ dando lugar al modelo *lineal.model.fit11*:

$$\begin{aligned} Price = & b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\ & b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\ & b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + \\ & b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P1^2 \end{aligned}$$

El rendimiento del modelo se expone en la siguiente tabla:

R^2	R^2 Ajustado	RMSE	RMSE 5-fold
0.34	0.3394	42928.75	43503.52

Cuadro 30: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P1$ elevado a 2

Comparando el error $RMSE$ medido sobre 5-fold , se aprecia que el rendimiento de este modelo es inferior al anterior.

Por su parte, la ejecución de las líneas de código correspondientes, nos devuelve la siguiente salida:

Call:

```
lm(formula = Price ~ . - H2p2 + I(P27p4^2) + I(H13p1^2) + I(H40p4^2) +
  I(P16p2^2) + I(P1^2) + I(P5p1^2), data = house)
```

Residuals:

Min	1Q	Median	3Q	Max
-178521	-19540	-6113	8676	454846

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.458e+05	2.595e+04	13.328	< 2e-16 ***
P1	5.975e-02	8.732e-03	6.842	7.99e-12 ***
P5p1	-1.538e+05	8.733e+04	-1.761	0.078335 .
P6p2	5.878e+04	1.712e+04	3.433	0.000598 ***
P11p4	7.827e+04	6.825e+03	11.468	< 2e-16 ***
P14p9	-1.278e+05	1.067e+04	-11.980	< 2e-16 ***
P15p1	-2.352e+05	3.225e+04	-7.294	3.12e-13 ***
P15p3	-2.057e+05	2.745e+04	-7.493	6.99e-14 ***
P16p2	-1.331e+05	5.834e+04	-2.281	0.022573 *
P18p2	-1.407e+06	6.112e+04	-23.028	< 2e-16 ***
P27p4	1.004e+06	2.594e+04	38.702	< 2e-16 ***
H8p2	-6.203e+04	1.840e+04	-3.370	0.000752 ***
H10p1	-4.185e+04	3.223e+03	-12.984	< 2e-16 ***
H13p1	-2.892e+05	7.524e+03	-38.434	< 2e-16 ***
H18pA	-5.088e+04	3.810e+03	-13.356	< 2e-16 ***
H40p4	6.471e+04	2.967e+03	21.811	< 2e-16 ***
I(P27p4^2)	-2.166e+06	7.148e+04	-30.303	< 2e-16 ***
I(H13p1^2)	2.892e+05	9.472e+03	30.533	< 2e-16 ***
I(H40p4^2)	-7.127e+04	2.908e+03	-24.511	< 2e-16 ***
I(P16p2^2)	2.610e+05	3.291e+04	7.933	2.23e-15 ***
I(P1^2)	-7.414e-09	1.538e-09	-4.820	1.44e-06 ***

```

I(P5p1^2)      6.663e+04  8.396e+04   0.794  0.427423
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42950 on 22762 degrees of freedom
Multiple R-squared:  0.34, Adjusted R-squared:  0.3394
F-statistic: 558.5 on 21 and 22762 DF,  p-value: < 2.2e-16

RMSE del modelo: 42928.16
Error RMSE del modelo lineal compuesto sobre 5-fold: 43503.53

```

En este extracto, se aprecia que el *p-value* asociado al test de Wall evaluado para el término que acabamos de añadir es significativo, por consiguiente, se considera que el coeficiente asociado a este término es igual a 0 y que, por lo tanto, el término puede ser eliminado.

Se decide volver a añadir un término con $P11p4$ elevado al cuadrado, dando lugar al modelo *lineal.model.fit12*:

$$\begin{aligned} Price = b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\ b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\ b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + \\ b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P1^2 + b_{22} * P11p4^2 \end{aligned}$$

El rendimiento del modelo se expone en la siguiente tabla:

R^2	R^2 Ajustado	RMSE	RMSE 5-fold
0.3403	0.3397	42920.07	43499.36

Cuadro 31: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P11p4$ elevado a 2

El rendimiento del modelo se ha mejorado, se prueba ahora añadir un término con $P6p2$ al cuadrado, dando lugar al modelo *lineal.model.fit13*:

$$\begin{aligned} Price = b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\ b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\ b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + \\ b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P1^2 + b_{22} * P11p4^2 + b_{23} * P6p2^2 \end{aligned}$$

El rendimiento del modelo se expone en la siguiente tabla:

R^2	R^2 Ajustado	RMSE	RMSE 5-fold
0.3403	0.3397	42919.2	43518.67

Cuadro 32: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P6p2$ elevado a 2

Se puede apreciar que, la capacidad de generalización de este modelo ha disminuído respecto del anterior. Por su parte, las sentencias para la generación de este modelo devolvieron la siguiente salida:

Call:

```
lm(formula = Price ~ . - H2p2 + I(P27p4^2) + I(H13p1^2) + I(H40p4^2) +
  I(P16p2^2) + I(P1^2) + I(P11p4^2) + I(P6p2^2), data = house)
```

Residuals:

Min	1Q	Median	3Q	Max
-179501	-19595	-6144	8741	454964

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.361e+05	1.683e+04	19.973	< 2e-16 ***
P1	6.115e-02	8.777e-03	6.967	3.34e-12 ***
P5p1	-8.506e+04	1.136e+04	-7.489	7.18e-14 ***
P6p2	6.182e+04	1.709e+04	3.617	0.000299 ***
P11p4	1.086e+05	1.248e+04	8.707	< 2e-16 ***
P14p9	-1.308e+05	1.078e+04	-12.130	< 2e-16 ***
P15p1	-2.492e+05	3.219e+04	-7.739	1.04e-14 ***
P15p3	-2.141e+05	2.761e+04	-7.754	9.30e-15 ***
P16p2	-1.406e+05	5.861e+04	-2.398	0.016479 *
P18p2	-1.408e+06	6.117e+04	-23.015	< 2e-16 ***
P27p4	1.003e+06	2.551e+04	39.313	< 2e-16 ***
H8p2	-7.011e+04	1.985e+04	-3.532	0.000413 ***
H10p1	-4.188e+04	3.224e+03	-12.991	< 2e-16 ***
H13p1	-2.892e+05	7.526e+03	-38.428	< 2e-16 ***
H18pA	-5.053e+04	3.812e+03	-13.256	< 2e-16 ***
H40p4	6.477e+04	2.969e+03	21.816	< 2e-16 ***
I(P27p4^2)	-2.169e+06	7.100e+04	-30.547	< 2e-16 ***
I(H13p1^2)	2.911e+05	9.485e+03	30.692	< 2e-16 ***
I(H40p4^2)	-7.148e+04	2.913e+03	-24.541	< 2e-16 ***

```

I(P16p2^2)    2.741e+05  3.331e+04   8.229 < 2e-16 ***
I(P1^2)       -7.614e-09  1.543e-09  -4.936 8.05e-07 ***
I(P11p4^2)   -5.748e+04  1.894e+04  -3.034 0.002412 **
I(P6p2^2)     7.786e+03  8.098e+03   0.961 0.336355
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```

Residual standard error: 42940 on 22761 degrees of freedom
 Multiple R-squared: 0.3403, Adjusted R-squared: 0.3397
 F-statistic: 533.7 on 22 and 22761 DF, p-value: < 2.2e-16

RMSE del modelo: 42919.2

Error RMSE del modelo lineal compuesto sobre 5-fold: 43518.67

En este extracto vuelve a ocurrir que, el *p-value* asociado al término que se acaba de añadir es significativo, lo cual nos lleva nuevamente a eliminar este término.

Se añade ahora un término con $H8p2$ elevado al cuadrado, dando lugar al modelo *lineal.model.fit14*:

$$Price = b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P1^2 + b_{22} * P11p4^2 + b_{23} * H8p2^2$$

El rendimiento del modelo se expone en la siguiente tabla:

R^2	R^2 Ajustado	RMSE	RMSE 5-fold
0.3405	0.3398	42914.32	43527.62

Cuadro 33: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H8p2$ elevado a 2

Se aprecia que el error *RMSE* ha descendido, pero el error *RMSE* medido sobre *5-fold* ha empeorado, por lo que se prefiere no añadir este término por el momento. Se prueba ahora a añadir un término con $P18p2$ al cuadrado en lugar del anterior, dando lugar al modelo *lineal.model.fit15*:

$$Price = b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P1^2 + b_{22} * P11p4^2 + b_{23} * P18p2^2$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.3408	0.3402	42901.84	43484.25

Cuadro 34: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P8p2$ elevado a 2

Se observa que el rendimiento del modelo ha mejorado. Se decide repetir el mismo proceso con $H10p1$ dando lugar al modelo *lineal.model.fit16*:

$$Price = b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P11p4^2 + b_{22} * P18p2^2 + b_{23} * H10p1^2$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.3558	0.3551	42413.21	42829.38

Cuadro 35: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H10p1$ elevado a 2

En este caso, el rendimiento del modelo ha sufrido una elevada mejora respecto de las anteriores interacciones.

Se repite este proceso con la variable $H18pA$ dando lugar al modelo *lineal.model.fit17*:

$$\begin{aligned}
Price = & b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\
& b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\
& b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + \\
& b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P11p4^2 + b_{22} * P18p2^2 + \\
& b_{23} * H10p1^2 + b_{24} * H18pA^2
\end{aligned}$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.3579	0.3572	42342.56	42744.66

Cuadro 36: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H18pA$ elevado a 2

El rendimiento del modelo mejora y se decide repetir el mismo proceso con la variable $H13p1$ dando lugar al modelo *lineal.model.fit18*:

$$\begin{aligned}
Price = & b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\
& b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\
& b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + \\
& b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P11p4^2 + b_{22} * P18p2^2 + \\
& b_{23} * H10p1^2 + b_{24} * H18pA^2 + b_{25} * H13p1^2
\end{aligned}$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.3579	0.3572	42342.56	42744.66

Cuadro 37: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $H13p1$ elevado a 2

El rendimiento del modelo obtenido es similar al del anterior modelo. Aunque el *p-value* asociado al término que se acaba de añadir no es significativo y se decide no añadir este término por el momento puesto que se comprueba que no mejora el rendimiento.

Se decide continuar con la variable $P14p9$ dando lugar al modelo *lineal.model.fit19*:

$$\begin{aligned}
Price = & b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\
& b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\
& b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + \\
& b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P11p4^2 + b_{22} * P18p2^2 + \\
& b_{23} * H10p1^2 + b_{24} * H18pA^2 + b_{25} * P14p9^2
\end{aligned}$$

El rendimiento del modelo se expone en la siguiente tabla:

R^2	R^2 Ajustado	RMSE	RMSE 5-fold
0.3583	0.3576	42330.3	42745.61

Cuadro 38: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con $P14p9$ elevado a 2

El rendimiento del modelo es muy similar al del anterior modelo. Para este modelo, el conjunto de comandos ejecutado devolvió la siguiente salida:

Call:

```
lm(formula = Price ~ . - H2p2 + I(P27p4^2) + I(H13p1^2) + I(H40p4^2) +
  I(P16p2^2) + I(P1^2) + I(P11p4^2) + I(P18p2^2) + I(H10p1^2) +
  I(H18pA^2) + I(H13p1^2) + I(P14p9^2), data = house)
```

Residuals:

Min	1Q	Median	3Q	Max
-196978	-19476	-5531	9561	453224

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.519e+05	1.699e+04	14.824	< 2e-16 ***
P1	4.238e-02	8.647e-03	4.901	9.60e-07 ***
P5p1	-1.048e+05	1.123e+04	-9.335	< 2e-16 ***
P6p2	9.266e+04	1.684e+04	5.503	3.77e-08 ***
P11p4	1.292e+05	1.353e+04	9.553	< 2e-16 ***
P14p9	-1.980e+05	2.417e+04	-8.191	2.72e-16 ***
P15p1	-2.407e+05	3.160e+04	-7.616	2.71e-14 ***
P15p3	-2.102e+05	2.715e+04	-7.740	1.03e-14 ***
P16p2	-1.111e+05	5.902e+04	-1.882	0.059806 .
P18p2	-1.711e+06	9.409e+04	-18.185	< 2e-16 ***

```

P27p4      9.761e+05  2.533e+04  38.542 < 2e-16 ***
H8p2      -9.606e+04  1.811e+04  -5.303 1.15e-07 ***
H10p1     2.970e+05  1.470e+04  20.205 < 2e-16 ***
H13p1     -2.959e+05  7.479e+03 -39.566 < 2e-16 ***
H18pA     -1.047e+05  6.952e+03 -15.055 < 2e-16 ***
H40p4      6.344e+04  2.957e+03  21.452 < 2e-16 ***
I(P27p4^2) -2.119e+06  7.054e+04 -30.039 < 2e-16 ***
I(H40p4^2) -6.903e+04  2.896e+03 -23.838 < 2e-16 ***
I(P16p2^2)  2.525e+05  3.402e+04  7.423 1.18e-13 ***
I(P1^2)     -5.261e-09  1.520e-09 -3.461 0.000539 ***
I(P11p4^2)  -1.084e+05  2.175e+04 -4.983 6.30e-07 ***
I(P18p2^2)  9.280e+06  1.966e+06  4.721 2.36e-06 ***
I(H10p1^2)  -2.570e+05  1.084e+04 -23.703 < 2e-16 ***
I(H18pA^2)  1.036e+05  1.203e+04  8.610 < 2e-16 ***
I(P14p9^2)  2.484e+05  6.842e+04  3.631 0.000283 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 42350 on 22758 degrees of freedom

Multiple R-squared: 0.3583, Adjusted R-squared: 0.3576

F-statistic: 508.3 on 25 and 22758 DF, p-value: < 2.2e-16

RMSE del modelo: 42330.3

Error RMSE del modelo lineal compuesto sobre 5-fold: 42745.61

Se aprecia que esta modificación, lleva a asociar un *p-value* al término con *P16p2*, puesto que no se desea desestimar este término del modelo, se decide no realizar esta modificación al modelo y no añadir un término con *P14p9* al cuadrado.

Se decide ahora, estudiar y aplicar otros términos no lineales con el objetivo de tratar de obtener mejoras más significativas que las obtenidas hasta ahora:

Del análisis exploratorio efectuado en la sección 2.1, se determinó que algunas variables presentaban distribuciones platicúrticas definidas en el intervalo [0,1] y muy centradas en el valor 0. Se piensa entonces que, añadiendo términos al modelo con estas variables aplicando transformaciones no lineales como raíces cuadradas que lleven a estas distribuciones a la normalidad, se pueda mejorar el rendimiento del modelo.

Se decide añadir un término con la raíz cuadrada de *P18p2* sin eliminar

el término cuadrático de $P18p2$, dando lugar al modelo *lineal.model.fit20*:

$$\begin{aligned} Price = & b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\ & b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\ & b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + \\ & b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P11p4^2 + b_{22} * P18p2^2 + \\ & b_{23} * H10p1^2 + b_{24} * H18pA^2 + b_{25} * \sqrt{P18p2} \end{aligned}$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.3633	0.3626	42165.14	42555.5

Cuadro 39: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con la raíz de $P12p2$

Se observa que el rendimiento del modelo ha mejorado, por lo que se decide continuar repitiendo este proceso con otras variables.

Se decide añadir otro término que considere la raíz de $H18pA$ dando lugar al modelo *lineal.model.fit21*:

$$\begin{aligned} Price = & b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\ & b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\ & b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + \\ & b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P11p4^2 + b_{22} * P18p2^2 + \\ & b_{23} * H10p1^2 + b_{24} * H18pA^2 + b_{25} * \sqrt{P18p2} + b_{26} * \sqrt{H18pA} \end{aligned}$$

El rendimiento del modelo se expone en la siguiente tabla:

R²	R² Ajustado	RMSE	RMSE 5-fold
0.3665	0.3658	42059.34	42446.75

Cuadro 40: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con la raíz de $H18pA$

El rendimiento del modelo vuelve a mejorar.

Se repite el proceso con la variable $P15p3$, dando lugar al modelo *lineal.model.fit22*:

$$\begin{aligned}
Price = & b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\
& b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\
& b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + b_{18} * H40p4^2 + \\
& b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P11p4^2 + b_{22} * P18p2^2 + \\
& b_{23} * H10p1^2 + b_{24} * H18pA^2 + b_{25} * \sqrt{P18p2} + b_{26} * \sqrt{H18pA} + b_{27} * \sqrt{P15p3}
\end{aligned}$$

R^2	R^2 Ajustado	RMSE	RMSE 5-fold
0.3673	0.3665	42032.55	42370.61

Cuadro 41: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con la raíz de $P15p3$

El rendimiento del modelo vuelve a mejorar.

Se repite el proceso con la variable $P1$ dando lugar al modelo *lineal.model.fit23*:

$$\begin{aligned}
Price = & b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\
& b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\
& b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + \\
& b_{18} * H40p4^2 + b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P11p4^2 + b_{22} * P18p2^2 + \\
& b_{23} * H10p1^2 + b_{24} * H18pA^2 + b_{25} * \sqrt{P18p2} + b_{26} * \sqrt{H18pA} + \\
& b_{27} * \sqrt{P15p3} + b_{28} * \sqrt{P1}
\end{aligned}$$

R^2	R^2 Ajustado	RMSE	RMSE 5-fold
0.3826	0.3818	41520.61	45273.47

Cuadro 42: Métricas de rendimiento evaluadas para el modelo compuesto que considera un término con la raíz de $P1$

El modelo obtenido muestra peor capacidad de generalización, además, el *p-value* asociado al término con $P15p3$ al cuadrado mostró un valor significativo, por lo que se decide no ejecutar esta iteración.

A continuación, puesto que la distribución de la variable $H18p4$ es fuertemente platicútica y asimétrica hacia la derecha (la mayor parte de la distribución

se halla muy concentrada y próxima al valor 1), se decide apostar por una transformación no lineal sobre la misma para hacerla tender a la normalidad e intentar mejorar los resultados. Aunque en anteriores interacciones ya se trató de efectuar esto añadiendo un término cuadrático, se decide probar con elevar la variable en este término a 4, dando lugar al modelo *lineal.model.fit23*

$$\begin{aligned} Price = & b_0 + b_1 * P1 + b_2 * P5p1 + b_3 * P6p2 + b_4 * P11p4 + b_5 * P14p9 + b_6 * P15p1 + \\ & b_7 * P15p3 + b_8 * P16p2 + b_9 * P18p2 + b_{10} * P27p4 + b_{11} * H8p2 + b_{12} * H10p1 + \\ & b_{13} * H13p1 + b_{14} * H18pA + b_{15} * H40p4 + b_{16} * P27p4^2 + b_{17} * H13p1^2 + \\ & b_{18} * H40p4^2 + b_{19} * P16p2^2 + b_{20} * P1^2 + b_{21} * P11p4^2 + b_{22} * P18p2^2 + \\ & b_{23} * H10p1^4 + b_{24} * H18pA^2 + b_{25} * \sqrt{P18p2} + b_{26} * \sqrt{H18pA} + b_{27} * \sqrt{P15p3} \end{aligned}$$

R^2	R^2 Ajustado	RMSE	RMSE 5-fold
0.3727	0.372	41851.26	42156.63

Cuadro 43: Métricas de rendimiento evaluadas para el modelo compuesto que sustituye el término cuadrático con $H10p1$ por otro término elevado a 4.

Se aprecia que el rendimiento de este modelo mejora en todas las métricas al modelo *lineal.model.fit22*.

Finalmente, se decide quedar como este modelo como el mejor modelo no lineal obtenido. Aunque la capacidad predictiva de este modelo sigue siendo reducida, resulta un modelo elaborado y complejo que trata de ajustarse de forma óptima a las características de las distribuciones del *dataset*.

2.3.3. Elaboración de modelos basados en k-NN

Por último, se elaboran y estudian modelos basados en *k-NN k Nearest Neighbour*. En el estudio de estos modelos se tratará de determinar el valor óptimo de *k* (número de vecinos más cercanos), las variables del *dataset* a considerar o transformaciones en las mismas.

Por su parte, para evaluar el rendimiento de estos modelos se emplearán las mismas métricas usadas en los anteriores epígrafes.

Primeramente, se realiza un primer modelo considerando sus parámetros por defecto (entre ellos $k=7$ y medida de distancia la distancia euclídea) y empleando todas las variables del *dataset*:

```
1 summary(knn.model.fit1)
```

```

2 cat('RMSE del modelo: ', evaluate.knn_rmse.house(knn.
    model.fit1), fill=T)
3 cat('Error RMSE del modelo kknn compuesto sobre 5-fold:'
    , mean(sapply(1:5, run_knn_k_fold_house,
4
        nombre, knn.model.fit1), fill=T))
5
6 # Reescalar P1 al intervalo [0,1]
7 min.p1 <- min(house$P1)

```

Script 13: Conjunto de sentencias para la elaboración y evaluación de un modelo 7-NN con sus parámetros por defecto y con todas las variables del *dataset*

El rendimiento de este modelo se muestra a continuación:

RMSE	RMSE 5-fold
22749.59	37751.58

Cuadro 44: Métricas de rendimiento evaluadas para el modelo 7-NN elaborado con todas las variables del *dataset* con todos sus parámetros.

La variable $P1$ del *dataset* se encuentra definida en el intervalo $[2, 7322564]$ mientras que el resto de variables independientes del *dataset* representan porcentajes y su dominio se halla comprendido en el intervalo $[0,1]$, por consiguiente, resulta conveniente reescalar esta variable al mismo intervalo que los porcentajes para evitar que predomine en el cálculo de las distancias:

```

1
2 house$P1_rescaled = (house$P1-min.p1)/(max.p1-min.p1)
3
4 # Entrenar otro modelo con P1 reescalado
5 knn.model.fit2 <- kknn(Price~.-P1, house, house)

```

Script 14:]Conjunto de sentencias para reescalar la variable $P1$ al intervalo $[0,1]$

Se entrena otro modelo 7-NN con parámetros por defecto y con la variable $P1$ reescalada obteniéndose los siguientes resultados:

RMSE	RMSE 5-fold
22749.59	38975.74

Cuadro 45: Métricas de rendimiento evaluadas para el modelo 7-NN elaborado con todas las variables del *dataset* con todos sus parámetros y $P1$ escalada al intervalo $[0,1]$.

Se aprecia que el rendimiento del modelo empeora respecto al modelo original que consideraba esta variable sin reescalar, por ello se decide investigar este hecho con más experimentos.

Por lo general, los atributos del *dataset* no se ajustan a distribuciones normales, se desea probar ahora a aplicar transformaciones en las variables independientes para tratar de hacerlas tender a distribuciones normales y comprobar si esto permite mejorar el rendimiento del modelo, dando lugar al modelo *knn.model.fit3*, para el cual, se considera la variable *P1* escalada al intervalo [0,1]:

```

1 # Entrenar otro modelo con transformaciones para
2   # normalizar las variables, pero sin reescalar P1
3 knn.model.fit4 <- kknn(Price~sqrt(P1)+P5p1+I(P6p2^(1/3))+
4   +sqrt(P11p4)+sqrt(P14p9)+
5   I(P15p1^2)+I(P15p3^(1/3))+I(
6   P16p2^2)+sqrt(P18p2)+sqrt(P27p4)+
```

Script 15: Conjunto de sentencias para realizar un modelo aplicando transformaciones en las variables independientes para tratar de hacerlas tender a la normalidad

Del anterior modelo se obtuvieron las siguientes métricas de rendimiento:

RMSE	RMSE 5-fold
21404.52	37803.76

Cuadro 46: Métricas de rendimiento evaluadas para el modelo 7-NN elaborado con todas las variables del *dataset* a las cuales se han aplicado transformaciones para hacerlas tender a la normalidad y la variable *P1* ha sido escalada al intervalo [0,1].

Se observa que el rendimiento del modelo ha mejorado notablemente, respecto del anterior modelo.

Se decide también, construir un modelo similar pero obviando la estandarización sobre la variable *P1*, dando lugar al modelo *knn.model.fit4* cuyos métricas se exponen a continuación:

RMSE	RMSE 5-fold
21404.52	35474.78

Cuadro 47: Métricas de rendimiento evaluadas para el modelo 7-NN elaborado con todas las variables del *dataset* a las cuales se han aplicado transformaciones para hacerlas tender a la normalidad. En esta ocasión, no se ha reescalado $P1$

Se observa por tanto, que el reescalado de la variable $P1$ lejos de mejorar los resultados lleva a empeorarlos, por lo que en adelante se decide considerar esta variable en su intervalo original.

Por último, se estudia el valor óptimo de k , para ello, se ejecuta una batería de experimentos haciendo el valor de k evaluando las métricas de rendimiento para cada valor obtenido.

Primeramente, la batería de experimentos se ejecutará considerando este último modelo elaborado (*knn.model.fit4*), obteniendo, para cada valor de k , los siguientes resultados:

k	RMSE	RMSE 5-fold
1	0	43964.86
3	12914.35	38590.6
7	21404.52	35474.78
13	25710.82	34625.77
21	28321.6	34538.84
51	31895	35136.77
75	33127.73	35587.12
103	34039.93	36011.39

Cuadro 48: Métricas de rendimiento evaluadas los modelos diferentes modelos generados para cada valor de k , a partir del modelo que aplicaba transformaciones a las variables del *dataset* para hacerlas tender a distribuciones normales. Se ha resaltado en negro las métricas óptimas.

Se observa que, aunque el modelo generado con $k=7$ ofrezca un error *RMSE* más bajo sobre el conjunto de entrenamiento, el modelo generado con $k=21$ ofrece un error más bajo medido sobre *5-fold* y, por consiguiente, ofrece mayor capacidad de generalización.

Por último, se decide repetir esta batería de experimentos tomando como referencia el modelo original que consideraba todas las variables in-

dependientes sin aplicar ningún tipo de transformación sobre las mismas (*knn.model.fit1*). Nuevamente, los resultados generados por este último se exponen en la siguiente tabla:

k	RMSE	RMSE 5-fold
1	0	46804.15
3	13695.25	41076.1
7	22749.59	37751.58
13	27198.15	36627.25
21	29788.8	36381.31
51	33361	36749.22
75	34578.15	37125.27
103	25468.19	37504.89

Cuadro 49: Métricas de rendimiento evaluadas para los diferentes modelos generados para cada valor de k , a partir del modelo que reescalaba la variable $P1$. Se ha resaltado en negro los valores óptimos.

Dadas los experimentos realizados, se puede tomar como modelo óptimo basado en k -NN es el que aplica transformaciones sobre todas las variables independientes y para un valor de $k=21$.

2.3.4. Comparación de modelos

Finalmente, se comparan los rendimientos del modelo de regresión lineal con los del modelo basado en k -NN y, adicionalmente, contra el modelo M5', el cual no se ha tratado en este proyecto y cuyas métricas de rendimiento han sido proporcionadas por el profesorado.

Para realizar esta comparativa, se tomarán una serie de métricas evaluadas para los distintos modelos sobre una serie de problemas. Estas métricas fueron sido recogidas anteriormente, y se hará uso de los correspondientes tests estadísticos para determinar si existen diferencias significativas entre los rendimientos de estos tres modelos considerados.

Puesto que se evalúan los rendimientos de modelos aplicados a los mismos problemas y, a priori, no se tiene información sobre la distribución que siguen los rendimientos de cada modelo, se hará uso de tests no paramétricos para muestras dependientes.

En primer lugar, se desea comparar los rendimientos del mejor modelo no

lineal con el mejor modelo basado en k -NN, por lo que se hará uso del test de Wilcoxon:

Se realiza esta comparación sobre las métricas evaluadas sobre los propios conjuntos de entrenamiento como sigue:

```

1 # Script para la comparación de modelos
2 results.train <- read.csv('./regr_train_alumnos.csv',
3   row.names = 1)
4 results.test <- read.csv('./regr_test_alumnos.csv', row.
5   names = 1)
6
7 # Sobre el conjunto de train
8 # Comparar out_train_lm con out_train_kknn (referencia)
9   con Wilcoxon
10 difs <- (results.train[,1] - results.train[,2]) /
11   results.train[,1]
12 wilc_1_2 <- cbind(ifelse(difs<0, abs(difs)+0.1, 0+0.1),
13   ifelse(difs>0, abs(difs)+0.1, 0+0.1))
14 colnames(wilc_1_2) <- c(colnames(results.train)[1],
15   colnames(results.train)[2])
16 head(wilc_1_2)
17
18 # Aplicar test y calcular R+ y R-
19 LMvsKNNTst <- wilcox.test(wilc_1_2[,1], wilc_1_2[,2],
20   alternative = "two.sided", paired=TRUE)
21 Rmas <- LMvsKNNTst$statistic
22 pvalue <- LMvsKNNTst$p.value
23
24 LMvsKNNTst <- wilcox.test(wilc_1_2[,2], wilc_1_2[,1],
25   alternative = "two.sided", paired=TRUE)
26 Rmenos <- LMvsKNNTst$statistic
27
28 cat('Test modelo lineal (R+) vs modelo k-NN:(R-)', fill=
29   T)
30 cat('Valor R+: ', Rmas, fill=T)
31 cat('Valor R-: ', Rmenos, fill=T)
32 cat('p-value del test: ', pvalue, fill=T)

```

Script 16: Conjunto de sentencias para aplicar el test de Wilcoxon tomando los rendimientos del conjunto de entrenamiento

Se obtuvo la siguiente salida:

```

Test modelo lineal (R+) vs modelo k-NN:(R-)
Valor R+: 10
Valor R-: 160
p-value del test: 0.000328064

```

Tomando un nivel de significancia de 0.05, el p -value nos lleva a rechazar la hipótesis nula y a admitir que **existen diferencias significativas entre**

los rendimientos de ambos modelos en el conjunto de train.

Se repite este análisis sobre las métricas evaluadas en test:

```
1 # Sobre el conjunto de test
2 # Comparar out_train_lm con out_train_kknn (referencia)
  con Wilcoxon
3 difs <- (results.test[,1] - results.test[,2]) / results.
  test[,1]
4 wilc_1_2 <- cbind(ifelse (difs<0, abs(difs)+0.1, 0+0.1),
  ifelse (difs>0, abs(difs)+0.1, 0+0.1))
5 colnames(wilc_1_2) <- c(colnames(results.test)[1],
  colnames(results.test)[2])
6 head(wilc_1_2)
7
8
9 # Aplicar test y calcular R+ y R-
10 LMvsKNNTst <- wilcox.test(wilc_1_2[,1], wilc_1_2[,2],
  alternative = "two.sided", paired=TRUE)
11 Rmas <- LMvsKNNTst$statistic
12 pvalue <- LMvsKNNTst$p.value
13
14 LMvsKNNTst <- wilcox.test(wilc_1_2[,2], wilc_1_2[,1],
  alternative = "two.sided", paired=TRUE)
15 Rmenos <- LMvsKNNTst$statistic
16
17 cat('Test modelo lineal (R+) vs modelo k-NN:(R-)', fill=
  T)
18 cat('Valor R+: ', Rmas, fill=T)
19 cat('Valor R-: ', Rmenos, fill=T)
20 cat('p-value del test: ', pvalue, fill=T)
```

Script 17: Conjunto de sentencias para aplicar el test de Wilcoxon tomando los rendimientos del conjunto de test

Se obtuvo la siguiente salida:

```
Test modelo lineal (R+) vs modelo k-NN:(R-)
Valor R+: 78
Valor R-: 93
p-value del test: 0.7660294
```

En este caso, el *p-value* asociado al test no nos permite rechazar la hipótesis nula y nos lleva a admitir que **no existen diferencias significativas en el rendimiento de los dos modelos sobre el conjunto de test.**

Por último, se desea comparar el rendimiento de los 3 modelos de forma simultánea y en caso de encontrar diferencias significativas, determinar qué

par de modelos son diferentes. Para este propósito comparamos los 3 modelos con el test de Friedman y hacemos uso del test *post-hoc* Holm para determinar qué par presenta diferencias significativas:

En primer lugar, comprobamos si existen diferencias significativas en las métricas evaluadas sobre el conjunto de *train*:

```
1 # Aplicar el test de Friedman sobre el conjunto de
  entrenamiento
2 test_friedman <- friedman.test(as.matrix(results.train))
3 test_friedman
```

Script 18: Conjunto de sentencias para aplicar el test de Friedman tomando las métricas sobre el conjunto de *train*

Friedman rank sum test

```
data: as.matrix(results.train)
Friedman chi-squared = 20.333, df = 2, p-value = 3.843e-05
```

Dado un valor de significancia de 0.05, el *p-value* nos lleva a rechazar la hipótesis nula y a afirmar que existen diferencias significativas en los rendimientos de los 3 modelos, para averiguar entre qué dos modelos existen diferencias significativas, haremos uso del test *post-hoc* de Holm:

```
1 # Aplicar el test post-hoc de Holm para averiguar qué
  par es diferente
2 tam <- dim(results.train)
3 groups <- rep(1:tam[2], each=tam[1])
4 pairwise.wilcox.test(as.matrix(results.train), groups, p
  .adjust = "holm", paired = TRUE)
```

Script 19: Conjunto de sentencias para aplicar el test post-hoc de Holm tomando las métricas sobre el conjunto de *train*

Pairwise comparisons using Wilcoxon signed rank test

```
data: as.matrix(results.train) and groups

  1      2
2 0.0031 -
3 0.0032 0.0032

P value adjustment method: holm
```

Considerando un nivel de significancia del 0.5, se observa que existen diferencias significativas en todos los modelos, es decir, no parece haber ningún modelo que ofrezca un rendimiento similar a cualquiera de los otros 2.

Repetimos este test para las métricas evaluadas sobre el conjunto de test:

```
1 # Aplicar el test de Friedman sobre el conjunto de test
2 test_friedman <- friedman.test(as.matrix(results.test))
3 test_friedman
```

Script 20: Conjunto de sentencias para aplicar el test Friedman tomando las métricas sobre el conjunto de test

Friedman rank sum test

```
data: as.matrix(results.test)
Friedman chi-squared = 8.4444, df = 2, p-value = 0.01467
```

Tomando nuevamente un nivel de significancia de 0.05, el *p-value* asociado al test nos lleva a rechazar la hipótesis nula y afirmar que existen diferencias significativas entre las variables.

Nuevamente aplicamos el test *post-hoc* Holm para encontrar los pares de variables entre los que existen diferencias significativas:

```
1 # Aplicar el test post-hoc de Holm para averiguar qué
  par es diferente
2 tam <- dim(results.test)
3 groups <- rep(1:tam[2], each=tam[1])
4 pairwise.wilcox.test(as.matrix(results.train), groups, p
  .adjust = "holm", paired = TRUE)
```

Script 21: Conjunto de sentencias para aplicar el test post-hoc de Holm tomando las métricas sobre el conjunto de test

Pairwise comparisons using Wilcoxon signed rank test

```
data: as.matrix(results.train) and groups

      1      2
2 0.0031 -
3 0.0032 0.0032
```

P value adjustment method: holm

Nuevamente, se aprecia que existen diferencias significativas en todos los modelos.

3. El dataset *vehicle*: Problema de clasificación.

3.1. Descripción del dataset *vehicle*

El dataset **vehicle**³ recoge valores medidos de diferentes características independientes de escala medidos sobre cada vehículo, obteniendo con ello, información sobre la Silueta del vehículo, información que será usada para determinar si el vehículo es uno de los siguientes modelos: Autobús de dos pisos, *Cheverolet van*, *Saab 9000* y un *Opel Manta 400*.

El dataset contiene un total de 846 instancias para las cuales se han tomado las siguientes 19 características⁴:

1. **Compactness**: Medida numérica entera de la compacidad del vehículo, calculada a partir del perímetro medio del vehículo y el área.
2. **Circularity**: Medida numérica entera de la circularidad del vehículo.
3. **Distance_circularity**: Medida numérica de la circularidad distante.
4. **Radius_rate**: Medida numérica entera del ratio del radio obtenida como $\frac{radiomayor - radiomenor}{radiomedio}$.
5. **Praxis_aspect_ratio**: Medida numérica entera del ratio de aspecto de praxis calculada como $\frac{ejemenor}{ejemayor}$.
6. **Max_length_aspect_ratio**: Medida numérica entera del ratio de aspecto de mayor longitud.
7. **Scatter_ratio**: Medida numérica entera de la relación de dispersión medida como $\frac{Inerciasobreel ejemenor}{inerciasobreel ejemayor}$.
8. **Elongatedness**: Medida numérica entera de la elongacidad del vehículo.
9. **Praxis_rectangular**: Medida numérica entera del área rectangular de la del vehículo calculada como $\frac{area}{longituddepraxis * anchuradelpraxis}$.

³Enlace al sitio web de Keel con información del dataset: <https://sci2s.ugr.es/keel/dataset.php?cod=68>

⁴Enlace al sitio web del que se extrajo originalmente el dataset: <http://archive.ics.uci.edu/ml/datasets/Statlog+Vehicle+Silhouettes>

10. **Length_rectangular**: Medida numérica entera de la longitud rectangular del vehículo.
11. **Major_variance**: Medida numérica entera de la mayor varianza.
12. **Minor_variance**: Medida numérica entera de la menor varianza.
13. **Giration_radius**: Medida numérica entera del ratio de giro.
14. **Major_skewness**: Medida numérica entera de la mayor asimetría del vehículo.
15. **Minor_skewness**: Medida numérica entera de la menor asimetría del vehículo.
16. **Minor_kurtosis**: Medida numérica entera del valor menor de kurtosis.
17. **Major_kurtosis**: Medida numérica entera del valor mayor de kurtosis.
18. **Hollows_ratio**: Medida numérica entera del ratio de *Hollow* del vehículo.

Por último, la variable dependiente del modelo es la siguiente:

- **Class**: Valor categórico que representa el modelo del vehículo a predecir: **van** (Chevrolet van), **saab** (Saab 9000), **bus** (Autobús de dos pisos) y **opel** (*Opel Manta 400*).

3.2. Análisis exploratorio del dataset *vehicle*

En primer lugar, se procede a realizar un análisis exploratorio de este *dataset*, con el objetivo de conocer información sobre las distribuciones de sus variables, las relaciones que pudieran existir entre las variables y conocer cualquier información de utilidad implícita en los datos.

Se comienza analizando la existencia de valores perdidos *Missing values* en el *dataset* y la distribución de los mismos. Para analizar si existe algún valor perdido realizamos lo siguiente:

```
1 # Comprobar la existencia de Missing Values
2 any(is.na(vehicle))
```

Script 22: Conjunto de sentencias para analizar la existencia de valores perdidos en el *dataset*

```
[1] FALSE
```

Se concluye, en este caso, que este *dataset* tampoco presenta valores perdidos.

3.2.1. Análisis de las variables del dataset

En esta sección, se llevará a cabo un análisis de las distribuciones de todas las variables del *dataset*.

Para analizar las distribuciones, nuevamente se hará uso de los estadísticos de posición (valor mínimo, valor máximo, primer cuartil, mediana, media aritmética y tercer cuartil, mientras que, para los atributos categóricos se hará uso de la moda), estadísticos de dispersión (desviación estándar) y los coeficientes de asimetría y de *kurtosis* como medidas para verificar la tendencia de la distribución a la distribución Normal. Además, se harán uso de los gráficos y análisis pertinentes para conocer, de forma visual, la forma de la distribución.

- **Compactness:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Compactness
Valor mínimo	73
Primer cuantil	87
Mediana	93
Media	93.68
Tercer cuantil	100
Valor máximo	119
Desviación estándar	8.234474
Coeficiente de asimetría	0.3805943
Coeficiente de Kurtosis	2.4607992

Cuadro 50: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *Compactness*

El dominio de esta variable se halla definido en el intervalo [73, 119]. La distribución de esta variable se halla centrada en su media (levemente desplazada a la izquierda) y, evaluando los estadísticos y el coeficiente de *Kurtosis*, podemos observar que los datos presentan cierta dispersión respecto del centro de la distribución.

Representamos un diagrama *boxplot* para visualizar de forma gráfica la posición de los datos a lo largo de la distribución y un histograma para conocer la forma de la distribución:

```
1 # Compactness
2 # Diagrama boxplot
3 graf.compactness.boxplot <- ggplot(vehicle, aes(x='
4   Compactness', y=Compactness)) +
5   geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
6     outlier.shape=8) +
7   labs(cation='center', y=' ', x=' ')
8 graf.compactness.boxplot
9
10 # Histograma
11 graf.compactness.histogram <- ggplot(vehicle, aes(
12   Compactness, fill=Class)) +
13   geom_histogram(binwidth=1, color='black', alpha=0.4) +
14   labs(y='frecuencia absoluta')
15 graf.compactness.histogram
```

Script 23: Conjunto de sentencias para dibujar y visualizar un diagrama *boxplot* y un histograma de la variable *Compactness*

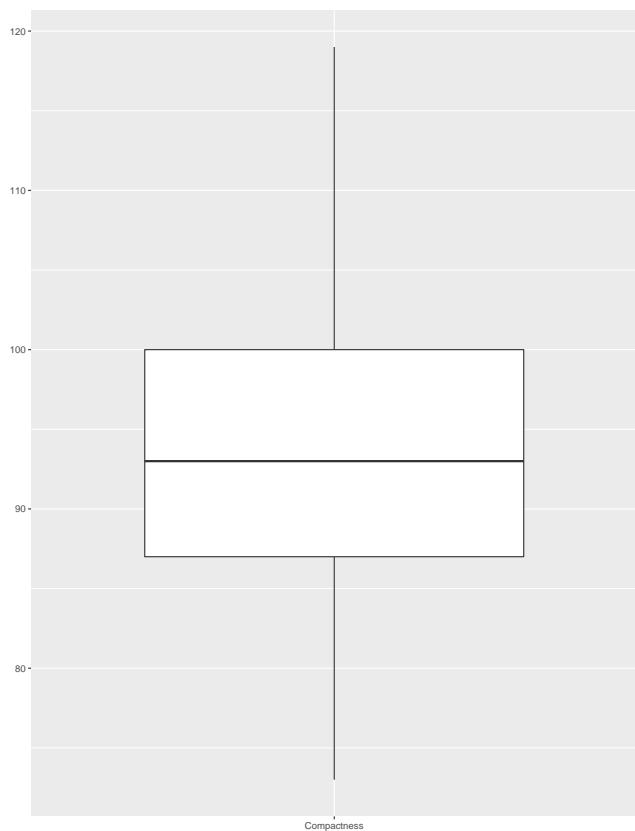


Figura 38: Histograma de la variable *Compactness*

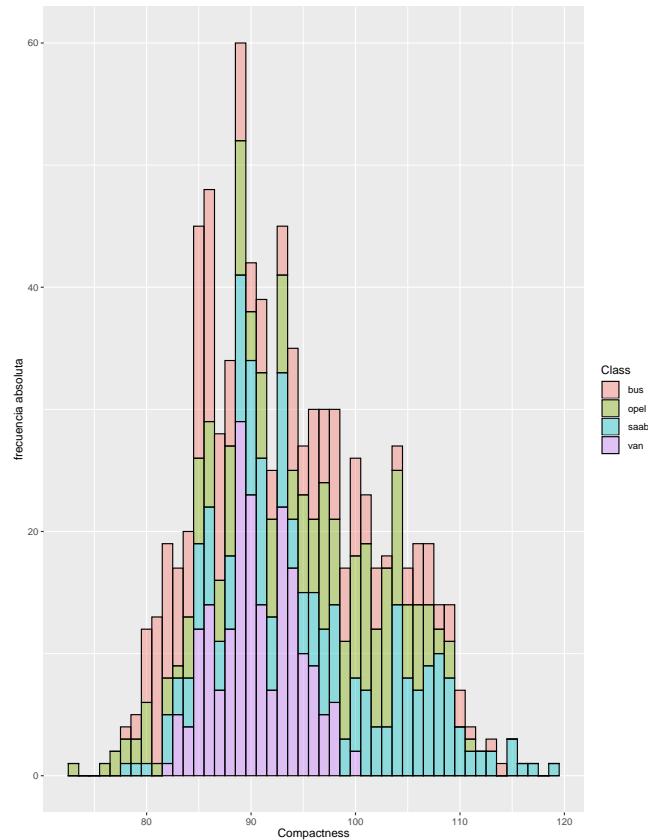


Figura 39: Histograma de la variable *Compactness*, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

Pese a existir cierta dispersión de los datos respecto de su centro, en el diagrama *boxplot* no se aprecian *outliers*.

Se puede apreciar a simple vista que, la distribución de la variable se aproxima a una distribución Normal, no obstante, conviene estudiar este aspecto de forma analítica, para lo cual haremos uso del test de *Shapiro-Wilk*:

```
1 # Estudiamos normalidad con el test de Shapiro-Milk
2 shapiro.test(vehicle$Compactness)
```

Script 24: Conjunto de sentencias para el test de Shapiro-Wilk sobre la distribución de *Compactness*

Shapiro-Wilk normality test

```

data: vehicle$Compactness
W = 0.97712, p-value = 2.99e-10

```

El *p-value* asociado al test no es significativo, lo que nos lleva a rechazar la hipótesis nula y a **negar que la distribución sea Normal.**

	H13p1
Valor mínimo	73
Primer cuantil	87
Mediana	93
Media	93.68
Tercer cuantil	100
Valor máximo	119
Desviación estándar	8.234474
Coeficiente de asimetría	0.3805943
Coeficiente de Kurtosis	2.4607992

Cuadro 51: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *Compactness*

- **Circularity:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Circularity
Valor mínimo	33.00
Primer cuantil	40.00
Mediana	44.00
Media	44.86
Tercer cuantil	49.00
Valor máximo	59.00
Desviación estándar	6.16986560602208
Coeficiente de asimetría	0.262332593782417
Coeficiente de Kurtosis	2.07338526143374

Cuadro 52: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *Circularity*

El rango de esta variable se halla definida en el intervalo [33, 59]. Observamos que la distribución de esta variable se halla levemente desplazada

a la izquierda de la distribución presentando cierta dispersión de los datos respecto de su centro de distribución.

Nuevamente, representamos un diagrama *boxplot* para visualizar de forma gráfica la organización de los datos y un histograma para visualizar la forma de la distribución:

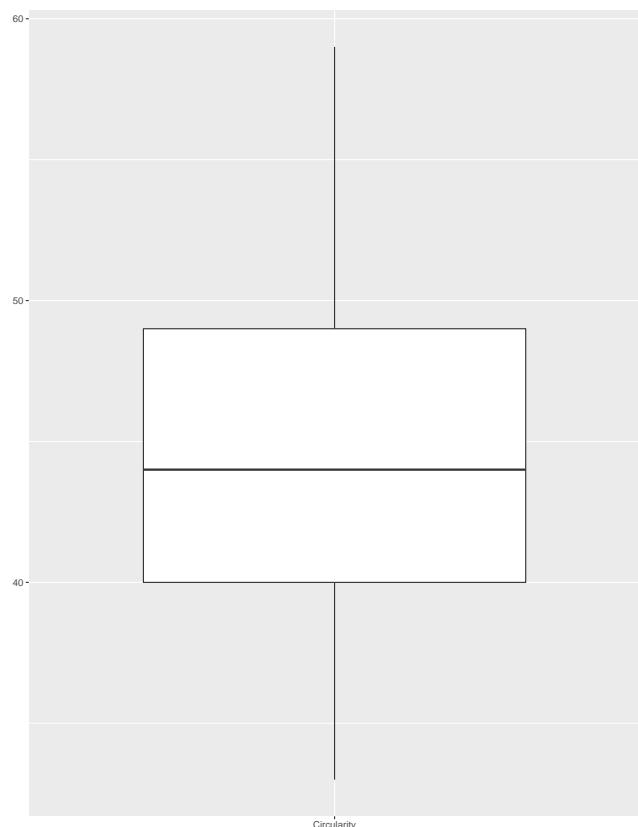


Figura 40: Diagrama *boxplot* de la variable *Circularity*

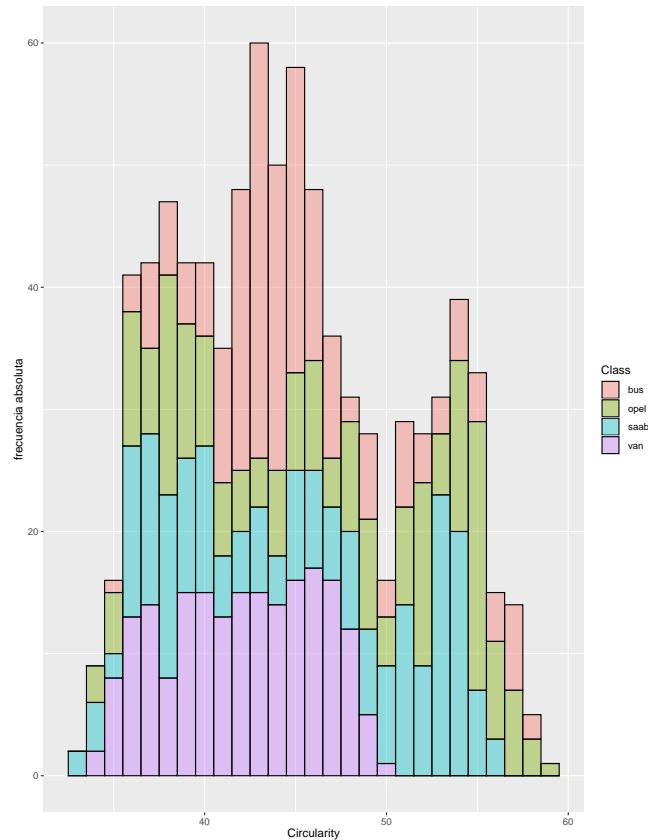


Figura 41: Histograma de la variable *Circularity*, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

Gráficamente, observamos que la forma de la distribución se asemeja a la distribución Normal, pero comprobamos este hecho aplicando el test de *Shapiro-Wilk* sobre esta variable.

```
1 # Estudiamos normalidad con el test de Shapiro-Wilk
2 shapiro.test(vehicle$Circularity)
```

Script 25: Conjunto de sentencias para el test de Shapiro-Wilk sobre la distribución de *Circularity*

Shapiro-Wilk normality test

```
data: vehicle$Circularity
W = 0.96404, p-value = 1.415e-13
```

El *p-value* asociado al test no es significativo, lo que nos lleva a rechazar la hipótesis nula y a **negar que la distribución sea Normal**.

- **Distance_circularity:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Distance_circularity
Valor mínimo	40.00
Primer cuantil	70.00
Mediana	80.00
Media	82.09
Tercer cuantil	98.00
Valor máximo	112.00
Desviación estándar	15.7715327049251
Coeficiente de asimetría	0.10703038451954
Coeficiente de Kurtosis	2.02021824392106

Cuadro 53: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *Distance_circularity*

La variable *Distance_circularity* se halla definida en el intervalo [40,112]. Se aprecia que el centro de la distribución se halla ligeramente desplazado a la izquierda presentando cierta dispersión de los datos respecto de este centro.

Nuevamente representamos un diagrama *boxplot* y un histograma para estudiar de forma gráfica la distribución de la variable:

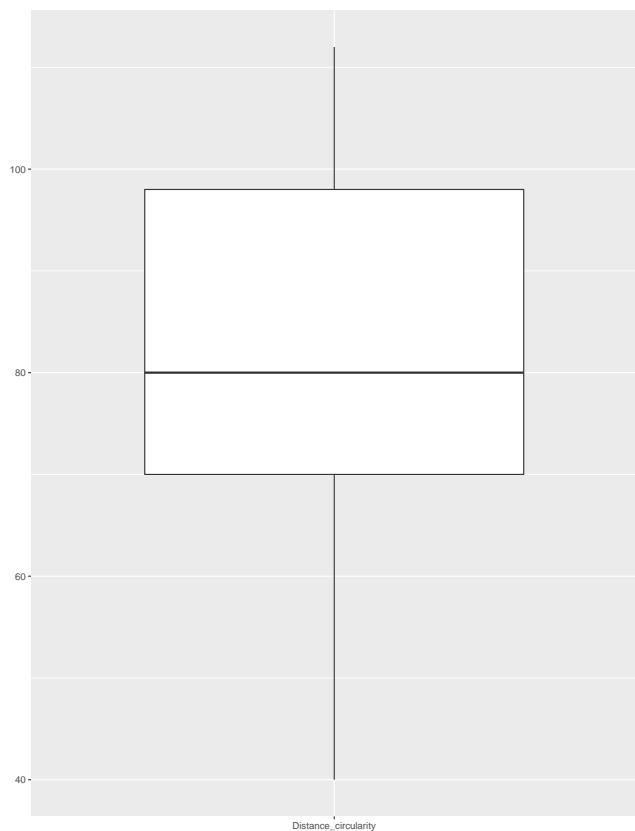


Figura 42: Diagrama *boxplot* de la variable *Distance_circularity*

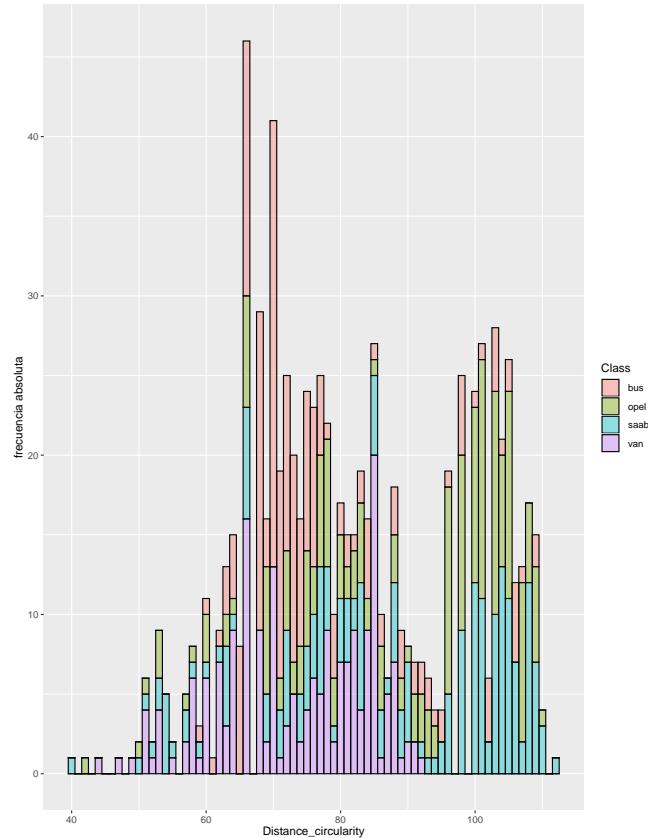


Figura 43: Histograma de la variable *Distance_circularity*, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

Si bien el centro de la distribución se halla ligeramente desplazado a la izquierda, a la derecha de este centro se sitúa una región muy densa. Nuevamente, se aprecia que la forma de la distribución se asemeja a una distribución Normal, no obstante, aplicamos de nuevo el test de *Shapiro-Wilk* para comprobarlo analíticamente.

```
1 # Estudiamos normalidad con el test de Shapiro-Wilk
2 shapiro.test(vehicle$Distance_circularity)
```

Script 26: Conjunto de sentencias para el test de Shapiro-Wilk sobre la distribución de *Distance_circularity*

Shapiro-Wilk normality test

```
data: vehicle$Distance_circularity
W = 0.95792, p-value = 7.422e-15
```

El *p-value* asociado al test no es significativo, lo que nos lleva a rechazar la hipótesis nula y a **negar que la distribución sea Normal**.

- **Radius_ratio:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Radius_ratio
Valor mínimo	104.0
Primer cuantil	141.0
Mediana	167.0
Media	168.9
Tercer cuantil	195.0
Valor máximo	333.0
Desviación estándar	33.4721830092547
Coeficiente de asimetría	0.390013381394982
Coeficiente de Kurtosis	3.29296054158576

Cuadro 54: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *Radius_ratio*

La variable se halla definida en el intervalo [104,333]. El centro de la distribución se halla desplazada a la izquierda, asimismo, la distribución presenta cierta dispersión de los datos respecto de este centro de distribución.

Representamos de nuevo un diagrama *boxplot* y un histograma para visualizar la forma de esta distribución:

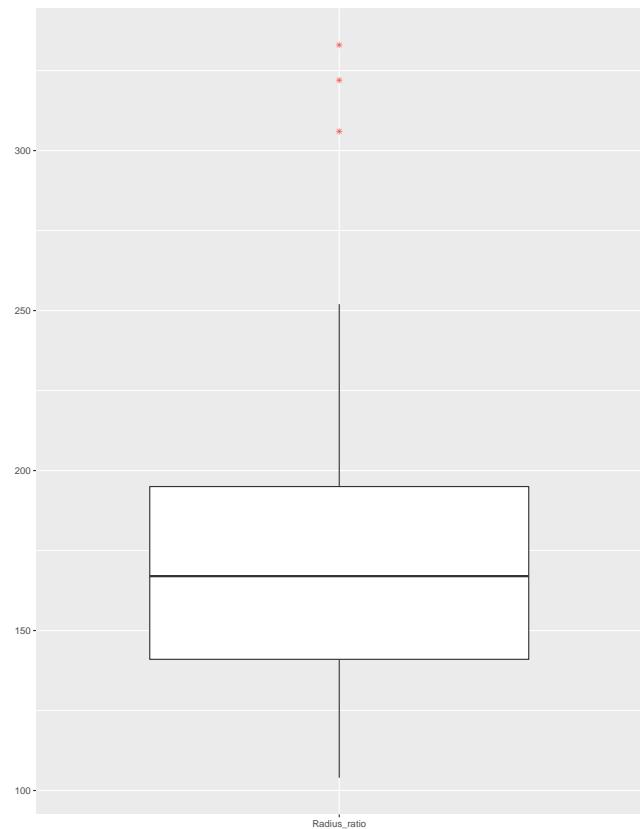


Figura 44: Diagrama *boxplot* de la variable *Radius_ratio*

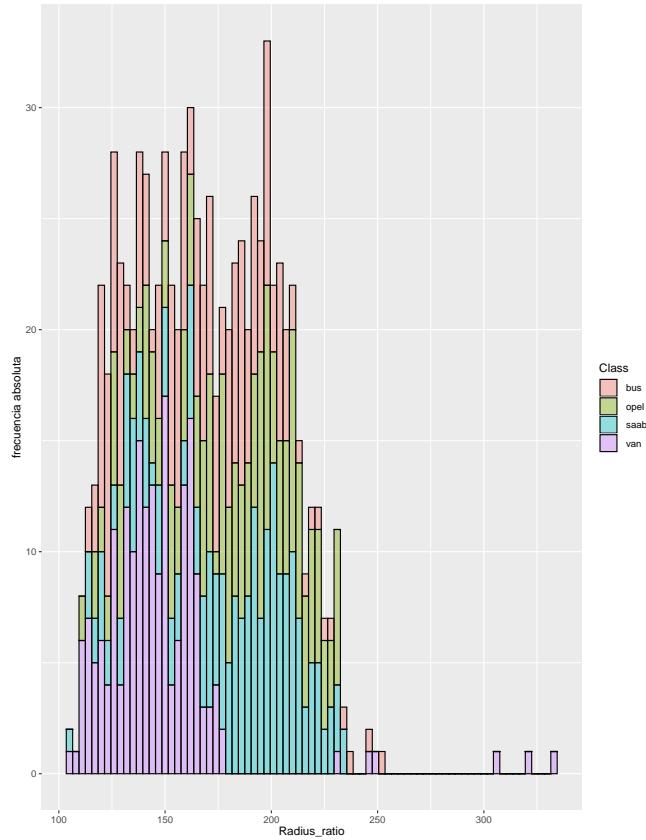


Figura 45: Histograma de la variable *Radius_ratio*, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

En las anteriores gráficas podemos observar como la mayor parte de la distribución se concentra a la izquierda de la media. Se observan algunos *outliers* muy puntuales situados muy alejados de la mayor parte de la distribución y próximos al valor máximo de la distribución, lo que despierta el interés por analizar esta situación:

Tomamos como referencia el intervalo de valores medido por el diagrama *boxplot* para la identificación de puntos pertenecientes a la distribución y exclusión de los *outliers*: $[Q1 - 1,5 * IRQ, Q3 + 1,5 * IRQ]$, donde $Q1$ es el primer cuartil, $Q3$ es el tercer cuartil e IRQ es el rango intercuartílico.

Calculamos este intervalo:

```

1 # Calcular rango de valores pertenecientes a la
   distribución
2 iqr.radious_ratio <- IQR(vehicle$Radius_ratio)
3 quantiles.radious_ratio <- quantile(vehicle$Radius_ratio
   )
4
5 q1.radious_ratio <- quantiles.radious_ratio[2]
6 q3.radious_ratio <- quantiles.radious_ratio[4]
7
8 cat('Intervalo de distribución: [', q1.radious_ratio-1.5
   *iqr.radious_ratio, ',',
   q3.radious_ratio+1.5*iqr.radious_ratio, ']', fill=T)
9

```

Script 27: Conjunto de sentencias para calcular el intervalo de valores que se consideran pertenecientes a la distribución

Intervalo de distribución: [60 , 276]

Los *outliers*, se sitúan por encima del límite superior de dicho intervalo, por lo que los aislamos y estudiamos más detenidamente:

```

1 # Aislamos los outliers que se encuentran por encima del
   límite superior
2 vehicle.outliers.radious_ratio <- vehicle %>% filter(
   Radius_ratio>276)
3 vehicle.outliers.radious_ratio %>% select(Radius_ratio,
   Class)

```

Script 28: Conjunto de sentencias para aislar los *outliers*

	Radius_ratio	Class
1	306	van
2	322	van
3	333	van

Los *outliers* corresponden a 3 vehículos del modelo *Chevrolet van* cuyas relaciones de radio son superiores a 300, valores superiores al del resto de vehículos de este mismo modelo. Para obtener más información sobre este hecho, convendría analizar el resto de variables y/o consultar documentación específica, por ello, por ahora se decide no tratar estos casos y dejarlos en nuestro *dataset*.

Aplicamos igualmente el test de *Shapiro-Wilk* para comprobar que esta distribución no tiende a una distribución Normal:

```

1 # Estudiamos normalidad con el test de Shapiro-Wilk
2 shapiro.test(vehicle$Radius_ratio)

```

Script 29: Conjunto de sentencias para el test de Shapiro-Wilk sobre la distribución de *Radius_ratio*

```

Shapiro-Wilk normality test

data: vehicle$Radius_ratio
W = 0.96983, p-value = 3.194e-12

```

El *p-value* asociado al test no es significativo, lo que nos lleva a rechazar la hipótesis nula y a **negar que la distribución sea Normal**.

- **Praxis_aspect_ratio:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Praxis_aspect_ratio
Valor mínimo	47
Primer cuantil	57
Mediana	61
Media	61.69
Tercer cuantil	65
Valor máximo	138
Desviación estándar	7.88825117206844
Coeficiente de asimetría	3.81478096027365
Coeficiente de Kurtosis	32.653110914701

Cuadro 55: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *Praxis_aspect_ratio*

La variable se halla definida en el intervalo [47, 138]. El centro de la distribución se halla profundamente desplazado a la izquierda, al tiempo que la distribución presenta una gran dispersión de los datos respecto de este centro de distribución.

Se aporta también un diagrama *boxplot* y un histograma para analizar la distribución de forma gráfica:

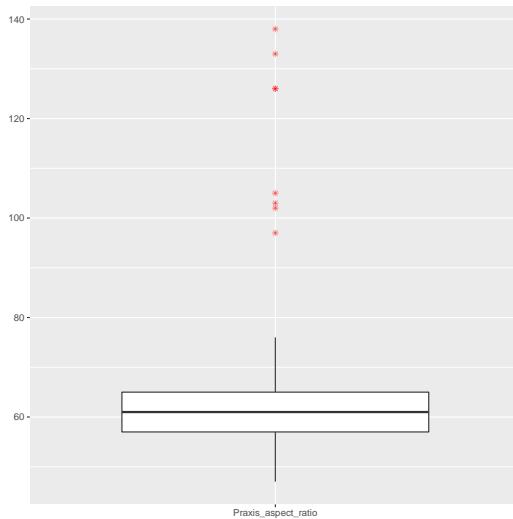


Figura 46: Diagrama *boxplot* de la variable Praxis_aspect_ratio

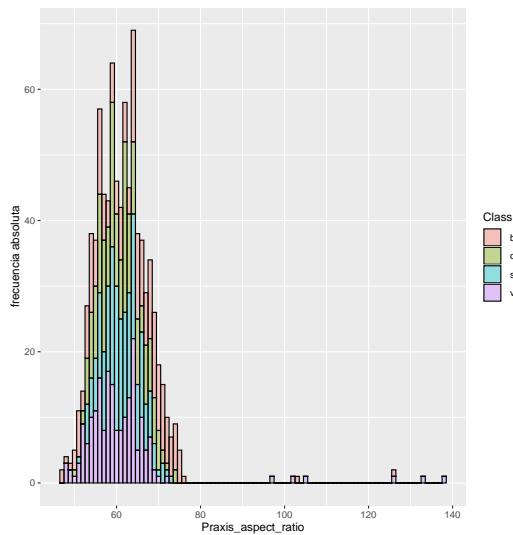


Figura 47: Histograma de la variable Praxis_aspect_ratio, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

Se aprecia a simple vista que la distribución no tiende a una distribución Normal.

- **Max_length_aspect_ratio:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Max_length_aspect_ratio
Valor mínimo	2
Primer cuantil	7
Mediana	8
Media	8.567
Tercer cuantil	10
Valor máximo	55
Desviación estándar	4.60121666113259
Coeficiente de asimetría	6.76636926656217
Coeficiente de Kurtosis	61.0239335012474

Cuadro 56: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *Max_length_aspect_ratio*

El intervalo de esta variable se halla definido en el intervalo [2,55]. Se aprecia que la distribución se halla muy desplazada a la izquierda de los datos y que su centro se halla próximo al valor mínimo de la distribución, por su parte, la distribución presenta gran dispersión de los datos respecto del centro de distribución, lo que llevaría a la aparición de *outliers* situados a la derecha de la distribución.

Analizamos gráficamente la distribución con un diagrama *boxplot* y con un histograma:

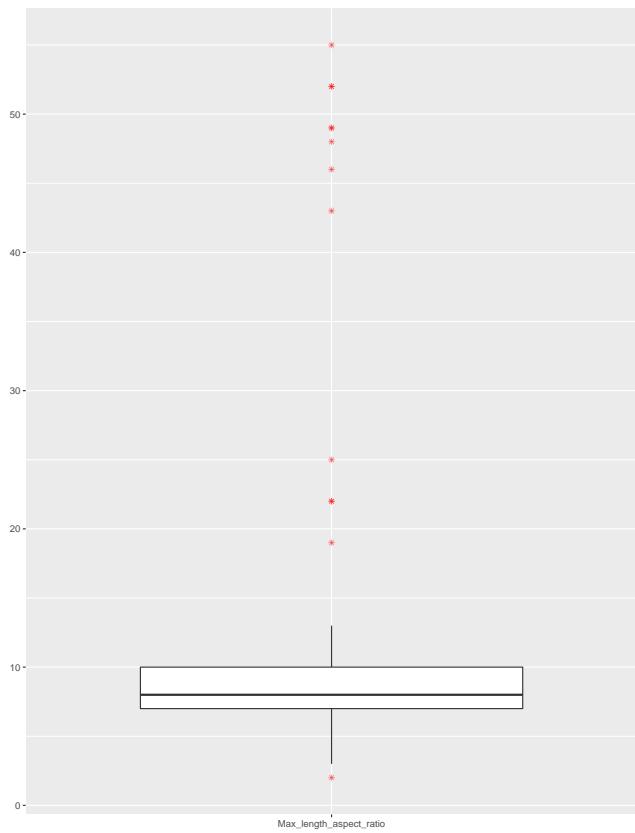


Figura 48: Diagrama *boxplot* de la variable *Max_length_aspect_ratio*

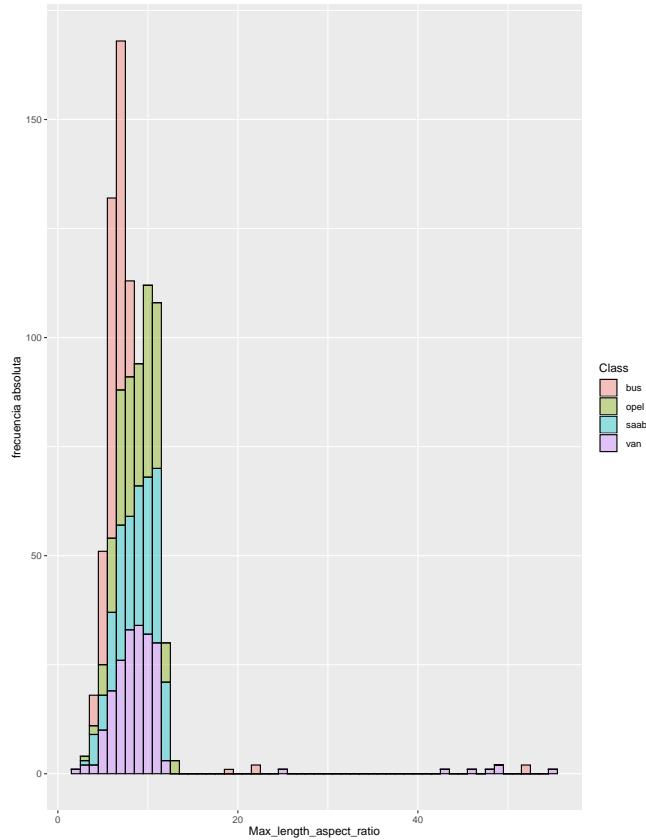


Figura 49: Histograma de la variable *Max_length_aspect_ratio*, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

Podemos comprobar en las anteriores gráficas, la existencia de instancias con valores de *Max_length_aspect_ratio* situados muy distantes del centro de la distribución. Estos *outliers* se corresponden con algunos vehículos de los modelos *Chevrolet Van* y autobuses de dos pisos.

Por último, se verifica a simple vista que la variable no se aproxima a una distribución Normal.

- **Scatter_ratio:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Scatter_ratio
Valor mínimo	112
Primer cuantil	146.2
Mediana	157
Media	168.8
Tercer cuantil	198
Valor máximo	265
Desviación estándar	33.2449780363009
Coeficiente de asimetría	0.604704397388578
Coeficiente de Kurtosis	2.38068214748892

Cuadro 57: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *Scatter_ratio*

El intervalo de esta variable se halla definido en el intervalo [112, 265]. Se observa que la distribución de esta variable se halla levemente desplazada a la izquierda presentando cierta dispersión de los datos respecto de su centro de distribución.

Se representa a continuación un diagrama *boxplot* y un histograma para visualizar la forma de la distribución:

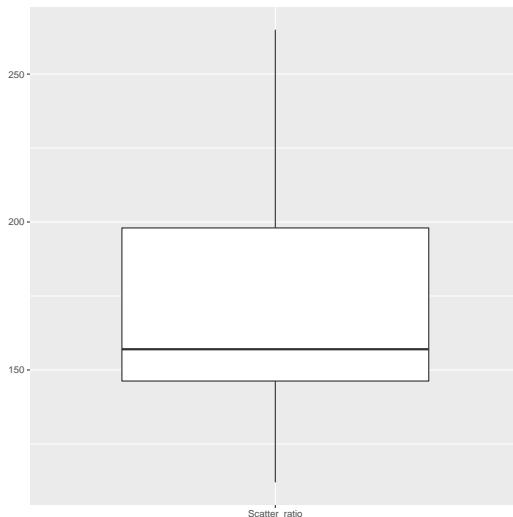


Figura 50: Diagrama *boxplot* de la variable *Scatter_ratio*

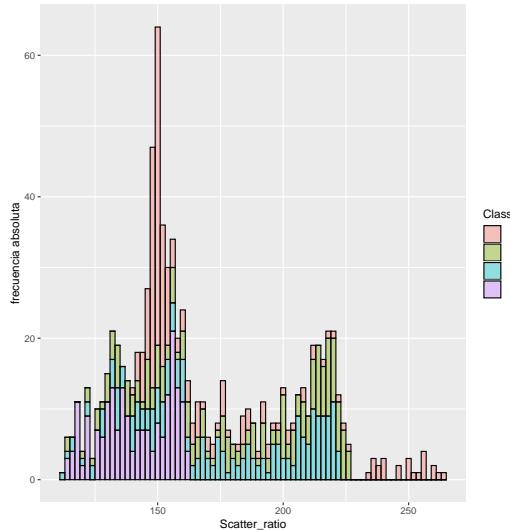


Figura 51: Histograma de la variable *Scatter_ratio*, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

En el anterior histograma, se verifica la existencia de autobuses de dos pisos con radios de dispersión mayores y diferenciados de todos los registros del *dataset*.

Por su parte, se verifica de forma visual que la distribución de la variable no es una distribución Normal.

- **Elongatedness:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Elongatedness
Valor mínimo	26
Primer cuantil	33
Mediana	43
Media	40.93
Tercer cuantil	46
Valor máximo	61
Desviación estándar	7.81155972001865
Coeficiente de asimetría	0.0477601827901232
Coeficiente de Kurtosis	2.13394641779649

Cuadro 58: Estadísticos de posición, de dispersión y coeficientes de asimetría y de *Kurtosis* calculados para el atributo *Elongatedness*

La variable que se halla definida en el intervalo [26,61] se halla ligeramente desplazada a la izquierda. El coeficiente de *Kurtosis* y los estadísticos de posición nos dan una idea de una dispersión de los datos respecto del centro de distribución.

Se representa a continuación un diagrama *boxplot* y un histograma con la distribución de esta variable:

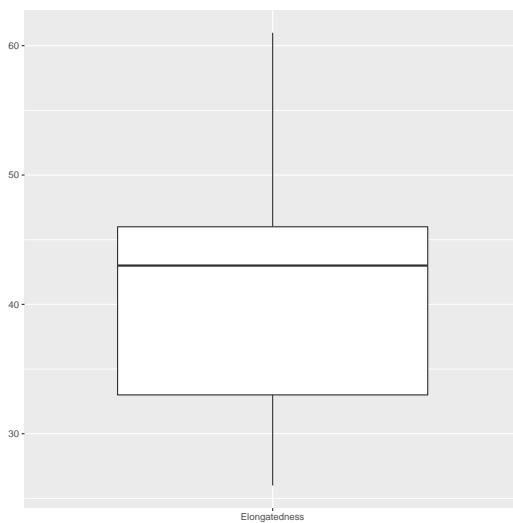


Figura 52: Diagrama *boxplot* de la variable *Elongatedness*

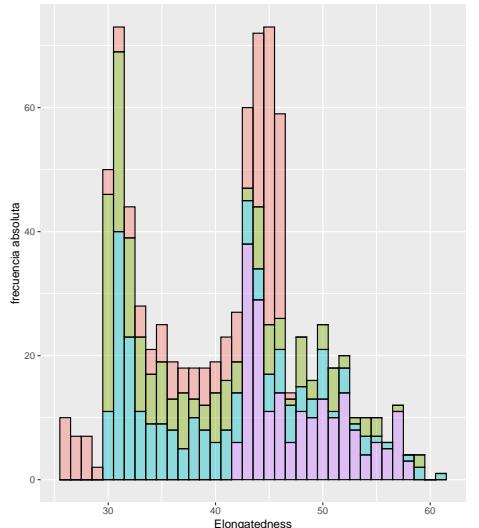


Figura 53: Histograma de la variable *Elongatedness*, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

Analizando gráficamente la distribución, se aprecia que la distribución presenta dos modas y que, no presenta una tendencia clara a una distribución Normal.

- **Praxis_rectangular:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Praxis_rectangular
Valor mínimo	17
Primer cuantil	19.00
Mediana	20.00
Media	20.58
Tercer cuantil	23.00
Valor máximo	29.00
Desviación estándar	2.59213832768472
Coeficiente de asimetría	0.769317320509767
Coeficiente de Kurtosis	2.60217533957887

Cuadro 59: Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo *Praxis_rectangular*

La variable se halla definida en el intervalo [17,29]. El centro de la distribución se halla desplazado a la izquierda de la distribución, por

su parte, la distribución presenta cierta dispersión de los datos respecto de su centro de distribución.

Analizamos la distribución gráficamente con ayuda de un diagrama *boxplot* y un histograma:

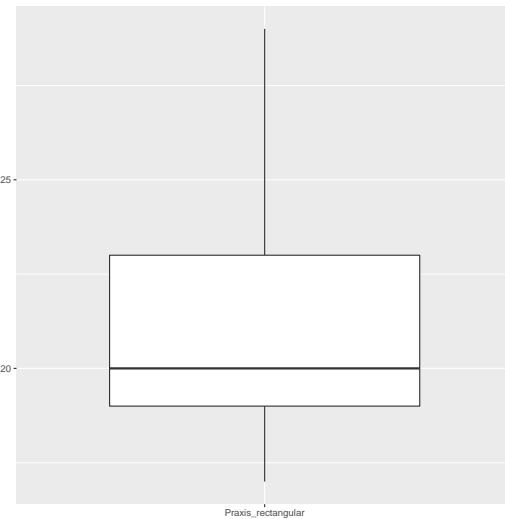


Figura 54: Diagrama *boxplot* de la variable *Praxis_rectangular*

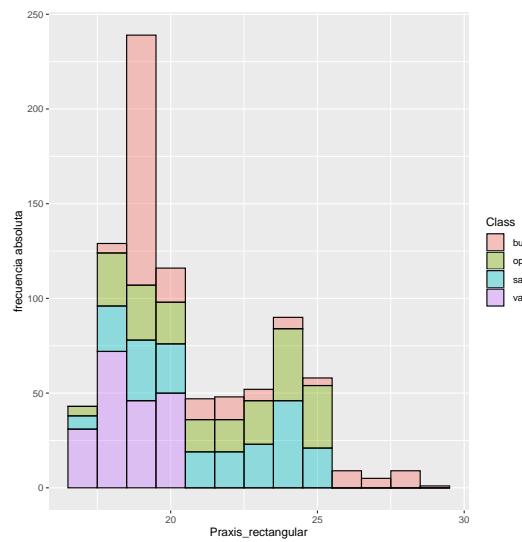


Figura 55: Histograma de la variable *Praxis_rectangular*, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

Se aprecia a simple vista que, la distribución de la variable difiera de la distribución Normal.

- **Length_rectangular:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Length_rectangular
Valor mínimo	118
Primer cuantil	137
Mediana	146
Media	148
Tercer cuantil	159
Valor máximo	188
Desviación estándar	14.515651573835
Coeficiente de asimetría	0.255904402558902
Coeficiente de Kurtosis	2.22736181882037

Cuadro 60: Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo Length_rectangular

La variable se halla definida en el intervalo [118, 188]. El centro de la distribución de halla desplazado a la izquierda de la distribución, al tiempo que los datos vuelven a presentar cierta dispersión respecto del centro.

Una vez más, volvemos a analizar la distribución de esta variable de forma gráfica con un diagrama *boxplot* y un histograma:

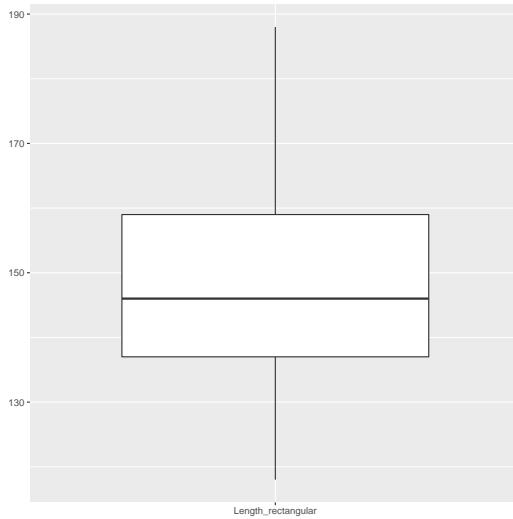


Figura 56: Diagrama *boxplot* de la variable Length_rectangular

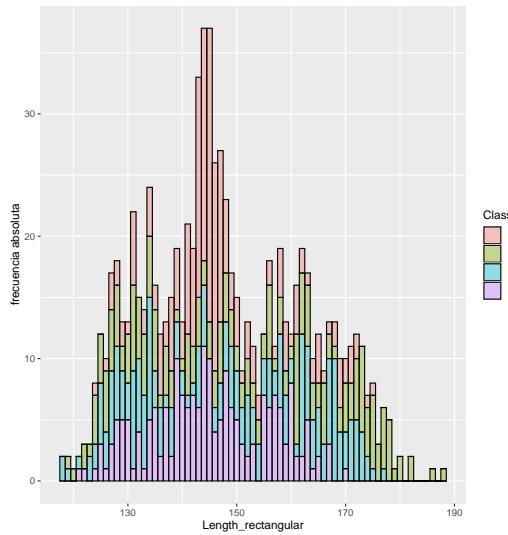


Figura 57: Histograma de la variable *Length_rectangular*, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

Gráficamente se observa que la distribución de la variable difiere de la distribución Normal

- **Major_variance:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Major_variance
Valor mínimo	130
Primer cuantil	167
Mediana	178.5
Media	188.6
Tercer cuantil	217
Valor máximo	320
Desviación estándar	31.3948365459496
Coeficiente de asimetría	0.650657596948011
Coeficiente de Kurtosis	3.11049171449922

Cuadro 61: Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo Major_variance

El rango de esta variable se sitúa en el intervalo [130,320]. Los estadísticos calculados nos dan una idea de una distribución asimétrica a la izquierda y que presenta cierta dispersión.

Representamos un diagrama *boxplot* y un histograma para conocer de forma visual la distribución de esta variable:

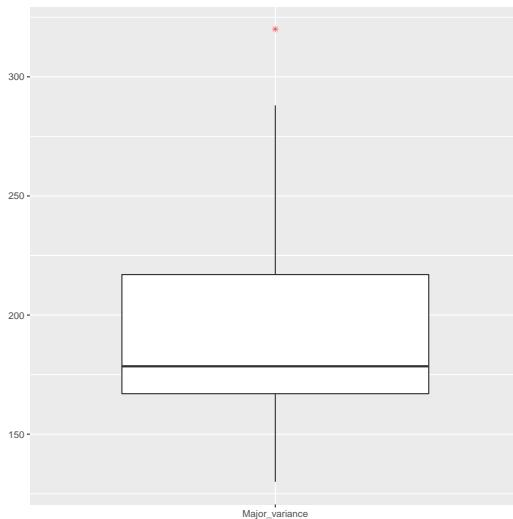


Figura 58: Diagrama boxplot de la variable Major_variance

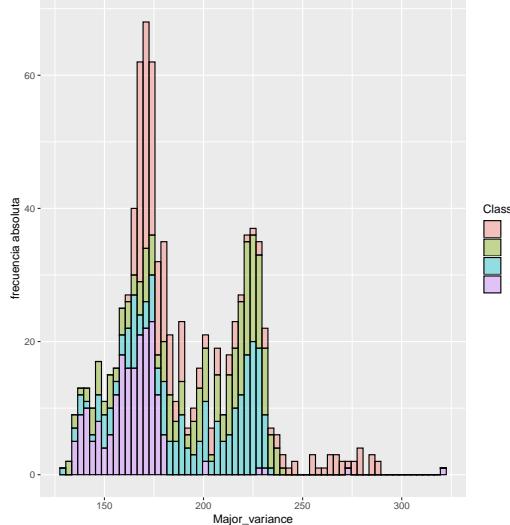


Figura 59: Histograma de la variable Major_variance, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

Gráficamente, podemos observar la presencia de algunos *outliers* puntuales y bien separados del resto de la distribución. Se decide analizar más detalladamente estos puntos:

```

1 # Analizar outliers por encima del valor 300
2 vehicle.outliers.major_variance <- vehicle %>% filter(
  Major_variance > 300)
3 vehicle.outliers.major_variance %>% select(Major_
  variance, Class)

```

Script 30: Conjunto de sentencias para aislar y mostrar los *outliers* de la variable *Major_variance*

	Major_variance	Class
1	320	van

Se trata por tanto, de un único vehículo del modelo *Chevrolet Van* que presenta un valor de varianza mayor muy superior del resto de vehículos. Al ser un único punto, se podría plantear la eliminación del mismo.

- **Minor_variance:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Minor_variance
Valor mínimo	184
Primer cuantil	318.2
Mediana	364.0
Media	439.9
Tercer cuantil	587
Valor máximo	1018
Desviación estándar	176.692613613023
Coeficiente de asimetría	0.834354120864896
Coeficiente de Kurtosis	2.7783360401455

Cuadro 62: Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo Minor_variance

La variable se halla definida en el intervalo [182, 1018]. Los estadísticos nos dan una idea de una distribución desplazada a la izquierda con dispersión de los datos respecto del centro de distribución.

Analizamos gráficamente la distribución de esta variable haciendo uso de un diagrama *boxplot* y un histograma:

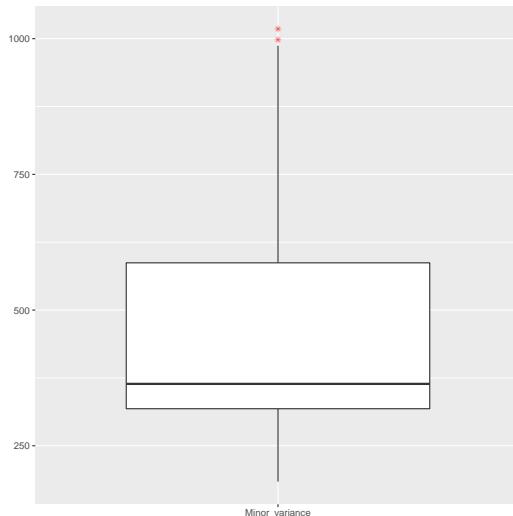


Figura 60: Diagrama *boxplot* de la variable Minor_variance

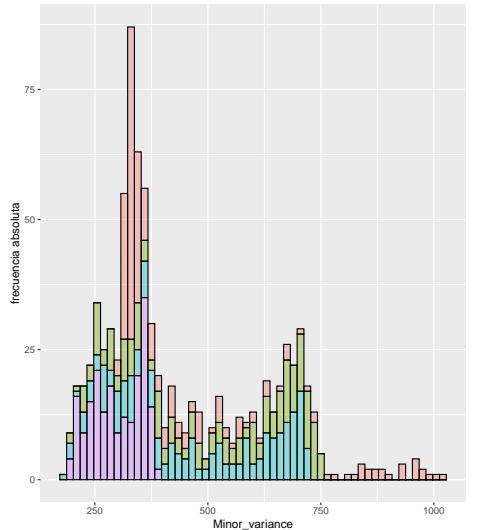


Figura 61: Histograma de la variable Minor_variance, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

Las gráficas identifican algunos *outliers* situados a la derecha de la distribución muy próximos al valor máximo de la misma. Se observa que, estos *outliers*, así como los valores más altos de esta variable se corresponden con autobuses de dos pisos, cuya varianza menor adopta estas medidas.

- **Gyration_radius:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Gyration_radius
Valor mínimo	109
Primer cuantil	149
Mediana	173
Media	174.7
Tercer cuantil	198
Valor máximo	268
Desviación estándar	32.5464898394173
Coeficiente de asimetría	0.279733433285242
Coeficiente de Kurtosis	2.50555984898307

Cuadro 63: Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo Gyration_radius

La variable se encuentra definida en el intervalo [109, 268]. La distribución presenta una asimetría a la izquierda y cierta dispersión de los datos respecto del centro de distribución.

Se representa, a continuación, un diagrama *boxplot* y un histograma para analizar gráficamente la distribución:

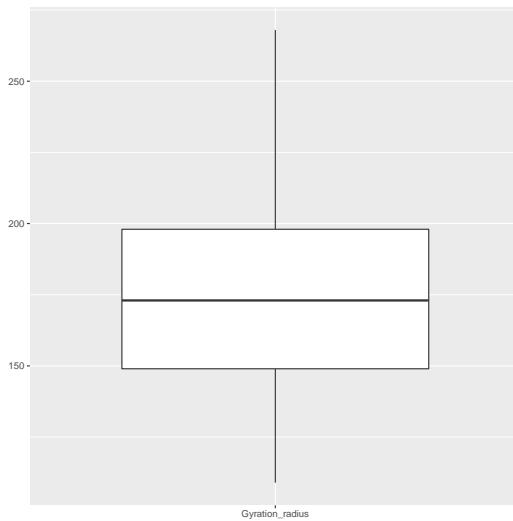


Figura 62: Diagrama boxplot de la variable Gyration_radius

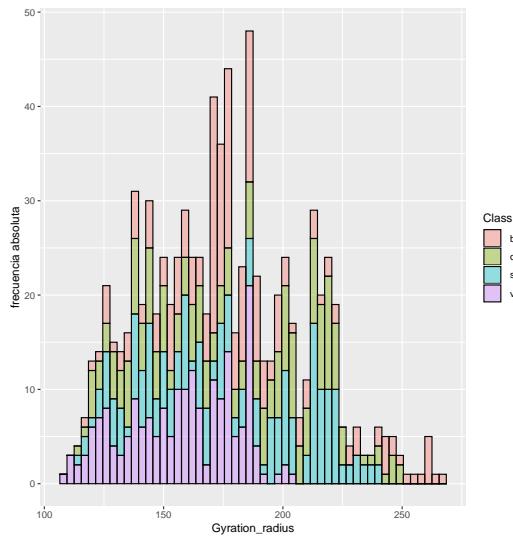


Figura 63: Histograma de la variable Gyration_radius, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

- **Major_skewness:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Major_skewness
Valor mínimo	59
Primer cuantil	67
Mediana	71.50
Media	72.46
Tercer cuantil	75
Valor máximo	135
Desviación estándar	7.48697406095933
Coeficiente de asimetría	2.06890653164845
Coeficiente de Kurtosis	14.298612692519

Cuadro 64: Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo Major_skewness

La variable se halla definida en el intervalo [59, 135]. El centro de la distribución se halla desplazado notablemente hacia la izquierda, asimismo, la distribución presenta gran dispersión de los datos respecto de este centro.

Para verificar más claramente estas características, se dibuja un gráfico *boxplot* y un histograma:

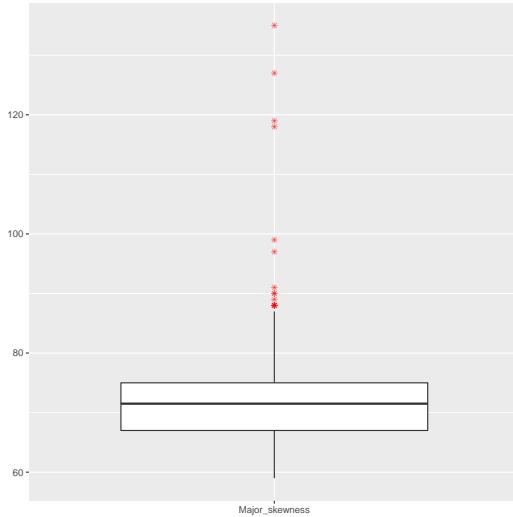


Figura 64: Diagrama boxplot de la variable Major_skewness

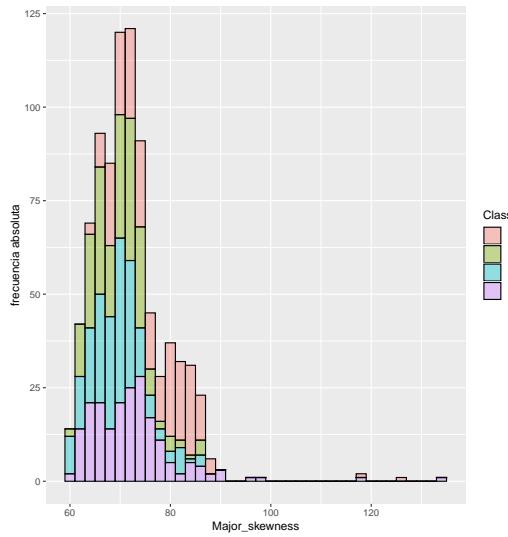


Figura 65: Histograma de la variable Major_skewness, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

Gráficamente, podemos apreciar la presencia de múltiples *outliers* situados a la derecha del centro de la distribución. Podemos, asimismo,

comprobar que se corresponden con autobuses o vehículos *Chevrolet Van* que presentan una gran asimetría en el eje mayor del vehículo.

- **Minor_skewness:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Minor_skewness
Valor mínimo	0
Primer cuantil	2
Mediana	6
Media	6.377
Tercer cuantil	9
Valor máximo	22
Desviación estándar	4.918352917405
Coeficiente de asimetría	0.772419256825095
Coeficiente de Kurtosis	3.08074291783783

Cuadro 65: Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo Minor_skewness

La variable *Minor_skewness* se halla definida en el intervalo [0,22] y su distribución presenta una asimetría a la izquierda de la distribución con dispersión de los datos respecto de su centro.

Nuevamente, visualizamos la distribución de esta variable haciendo uso de un diagrama *boxplot* y un histograma:

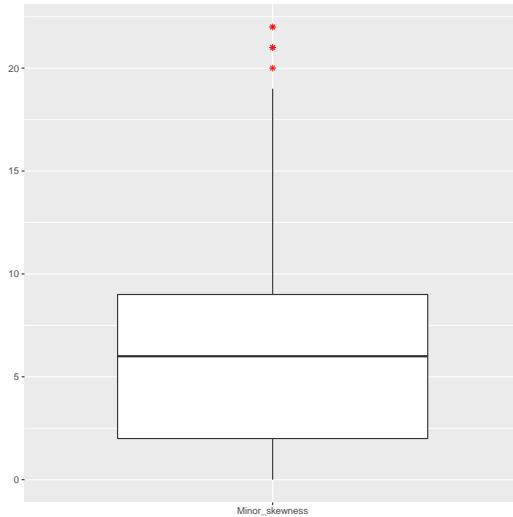


Figura 66: Diagrama boxplot de la variable Minor_skewness

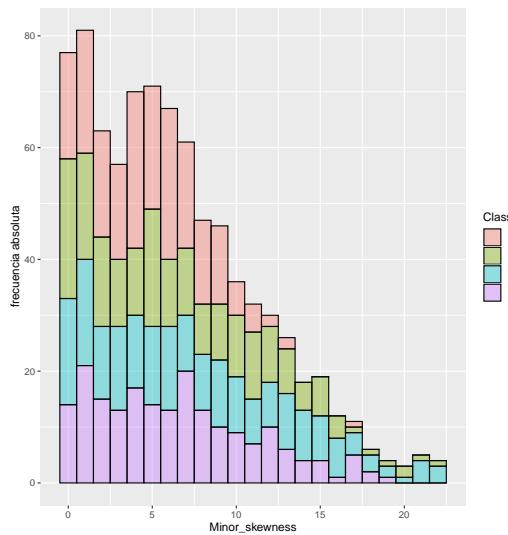


Figura 67: Histograma de la variable Minor_skewness, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

- **Minor_kurtosis:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Minor_kurtosis
Valor mínimo	0
Primer cuantil	5
Mediana	11
Media	12.6
Tercer cuantil	19
Valor máximo	41
Desviación estándar	8.93124026946145
Coeficiente de asimetría	0.688102630351292
Coeficiente de Kurtosis	2.85279054865626

Cuadro 66: Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo Minor_kurtosis

La variable se halla definida en el intervalo [0,41]. El centro de la distribución se halla desplazado a la izquierda de la distribución y dispersión de los datos respecto de su centro, lo que dará lugar a *outliers* especialmente situados a la derecha de la distribución.

Representamos gráficamente la distribución de la variable mediante un diagrama *boxplot* y un histograma:

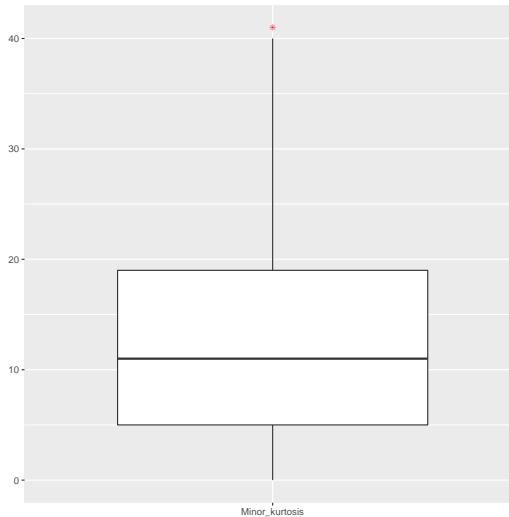


Figura 68: Diagrama boxplot de la variable Minor_kurtosis

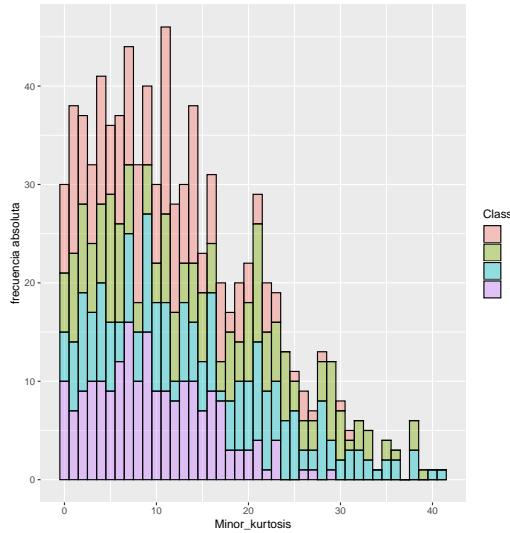


Figura 69: Histograma de la variable Minor_kurtosis, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

- **Major_kurtosis:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Major_kurtosis
Valor mínimo	176
Primer cuantil	184
Mediana	188
Media	188.9
Tercer cuantil	193
Valor máximo	206
Desviación estándar	6.16394935781931
Coeficiente de asimetría	0.248099903325358
Coeficiente de Kurtosis	2.40233260861461

Cuadro 67: Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo Major_kurtosis

La variable se halla definida en el intervalo [176, 206] y presenta cierta dispersión a la izquierda de la distribución con cierta dispersión de los datos respecto de su centro de distribución.

Representamos gráficamente esta información haciendo uso de un diagrama *boxplot* y un histograma:

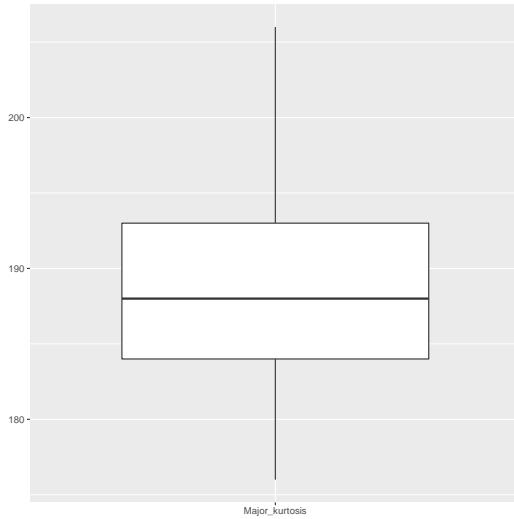


Figura 70: Diagrama boxplot de la variable Major_kurtosis

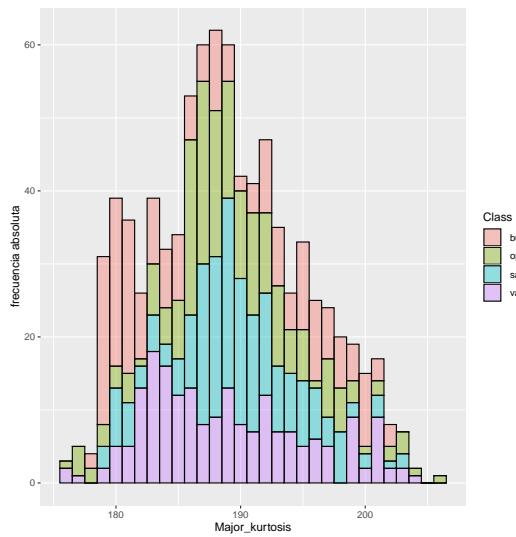


Figura 71: Histograma de la variable Major_kurtosis, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

- **Hollows_ratio:** Los estadísticos medidos sobre esta variable se resumen en la siguiente tabla:

	Hollows_ratio
Valor mínimo	181
Primer cuantil	190.2
Mediana	197
Media	195.6
Tercer cuantil	201
Valor máximo	211
Desviación estándar	7.43879742912235
Coeficiente de asimetría	-0.225939768186461
Coeficiente de Kurtosis	2.18428072350072

Cuadro 68: Estadísticos de posición, de dispersión y coeficientes de asimetría y de Kurtosis calculados para el atributo Hollows_ratio

La variable se encuentra definida en el intervalo [181, 211]. Al contrario que las otras variables, el centro de la distribución se halla desplazado a la derecha y la distribución presenta cierta dispersión de los datos respecto de su centro.

Nuevamente, analizamos de forma gráfica esta distribución haciendo uso de un diagrama *boxplot* y un histograma:

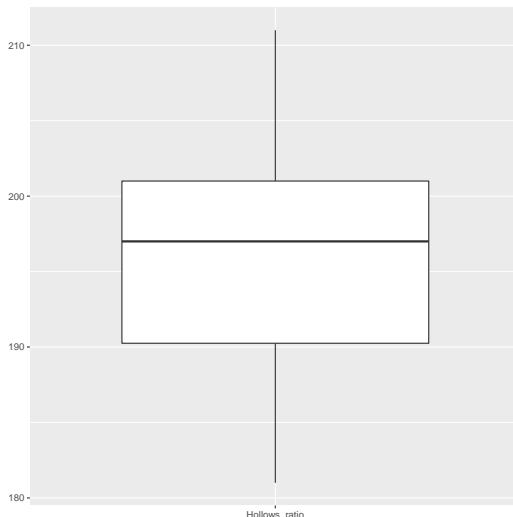


Figura 72: Diagrama boxplot de la variable Hollows_ratio

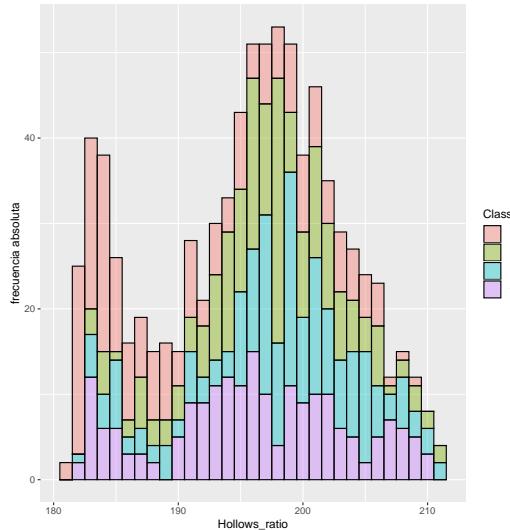


Figura 73: Histograma de la variable Hollows_ratio, se ha representado por colores las porciones del conjunto de datos que pertenece a cada clase

- **Class:** Finalmente, la variable dependiente del *dataset*, se trata de una variable categórica. La distribución de los valores se resume en la siguiente tabla:

	bus	opel	saab	van
Nº de instancias	218	212	217	199

Cuadro 69: Tabla de contingencia con la distribución de la variable *Class*

Igualmente, representamos estos datos de forma gráfica haciendo uso de un diagrama de barras:

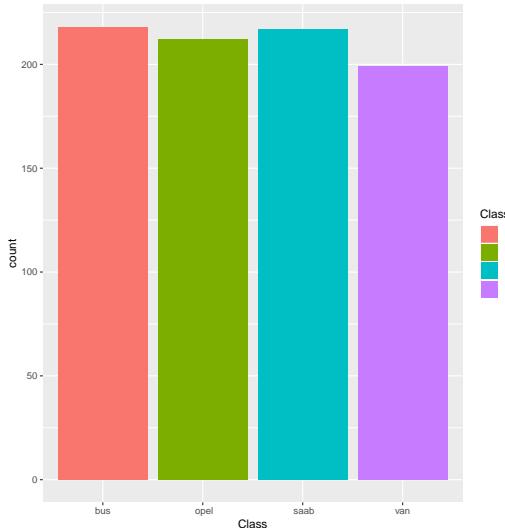


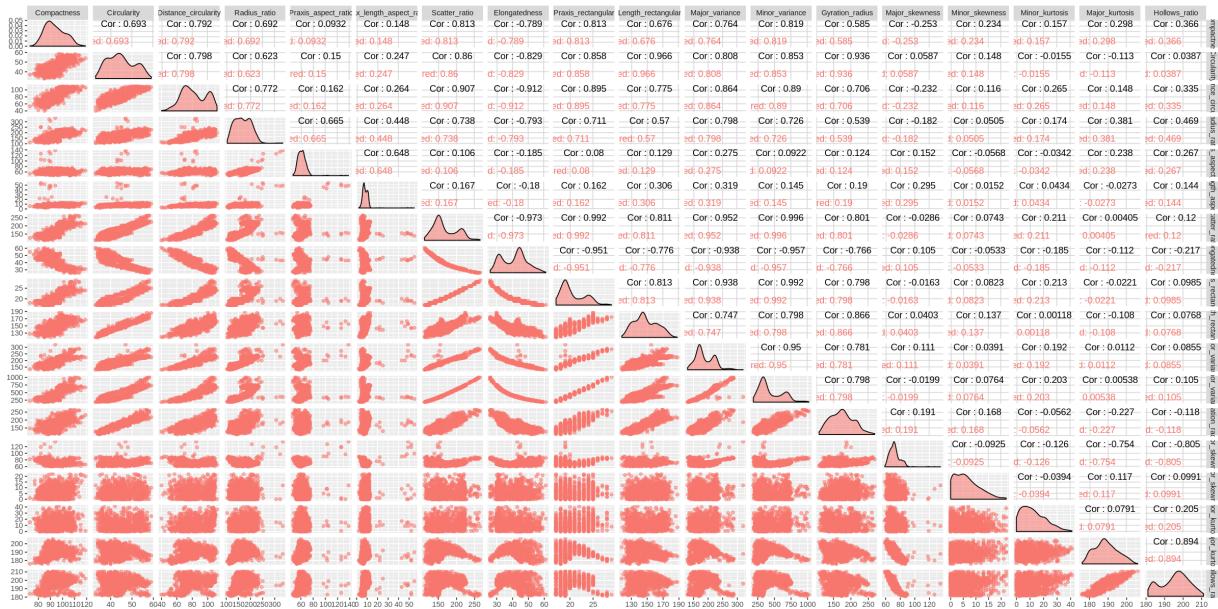
Figura 74: Diagrama de barras de la variable *Class*

Se aprecia que, las clases del *dataset* se encuentran más o menos balanceadas, lo que resulta adecuado para la construcción de modelos predictivos.

3.2.2. Análisis de las relaciones entre las variables

Una vez analizadas las distribuciones de las variables, se procederá a analizar las relaciones existentes entre las mismas, con la finalidad de determinar dependencias que existan entre las variables y que deban ser tenidas en cuenta para la elaboración de modelos.

Primeramente, se estudiarán las relaciones de forma gráfica haciendo uso de diagramas *scatterplot* con el fin de determinar a simple vista si las relaciones entre las variables se pueden aproximar a una función conocida y estudiar cómo varía una respecto a otra. Para esta representación se excluirá la variable *Class*:



En segundo lugar, se decide cuantificar de forma analítica las relaciones lineales entre las variables, para lo cual, puesto que muchas de las variables del *dataset* no tienden a la normalidad, ni tampoco en muchos casos, parecen mostrar una relación monotónica se usará como métrica el coeficiente de correlación de *Kendall*, el cual se muestra en la siguiente figura:

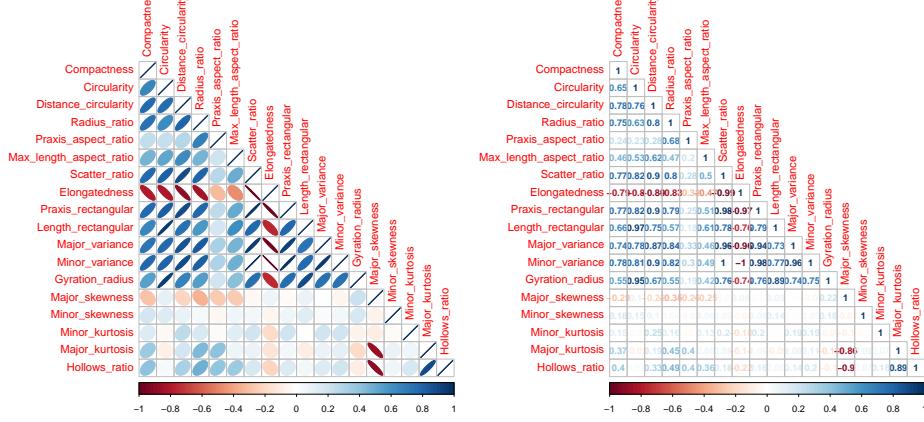


Figura 75: A la derecha se muestra una matriz triangular inferior donde cada celda muestra el valor del coeficiente de *Kendall* asociado a cada par de variables, nótese que un color cercano a azul oscuro indica una correlación lineal directa, mientras que un color rojo indica una correlación lineal inversa. A la izquierda se representa otro correlograma donde el valor del coeficiente ha sido sustituido por una esfera cuyo color y forma muestran el grado de correlación lineal entre las variables.

Puesto que muchas de las variables han sido calculadas a partir de cálculos matemáticos cuyos términos son compartidos en otras variables, resulta lógico que existe una fuerte correlación entre estas variables.

3.3. Elaboración de modelos predictivos

Una vez analizado el *dataset* y, conocidas sus principales características, se tratará de usar la información contenida en el mismo para la elaboración de modelos predictivos.

Para la elaboración de los modelos predictivos, se construirán modelos basados en 3 tipos diferentes: modelos basados en *k-NN* (en este caso, aplicado a clasificación), modelos basados en *LDA* (*Lineal Discriminant Analisys*) y

QDA (*Quadratic Discriminant Analysis*).

Al tratarse de problemas de clasificación, el error se medirá usando la métrica *CCR* (*Correct Classification Rate*). Este error se medirá sobre el conjunto de entrenamiento y, adicionalmente, para evaluar la capacidad de generalización de los modelos, se evaluará en un *5-fold*.

3.3.1. Elaboración de modelos basados en k-NN

Primeramente, se estudiarán y elaborarán modelos de clasificación basados en k-NN usando las métricas de rendimiento anteriormente descritas.

En primer lugar, se creará un modelo considerando todas las variables independientes del *dataset* y sin aplicar sobre las mismas ningún tipo de transformación, considerando un valor de $k=7$, dando lugar al modelo *knn.model.fit1*

```
1 # Modelo con k=7
2 knn.model.fit1 <- knn(train=vehicle[,-ncol(vehicle)],
3                         test=vehicle[,-ncol(vehicle)],
4                         cl=vehicle[,ncol(vehicle)], k=7)
5
6 # Evaluar error CCR train
7 cat('Accuracy sobre el conjunto de train: ', postResample
8     (knn.model.fit1,
9      vehicle$Class)[1],
10     fill=T)
11 cat('Accuracy sobre 10-fold: ', mean(sapply(1:10, run_knn_
12     k_fold_vehicle,
13                           fold_basename, 7)), fill
14     =T)
```

Script 31: Conjunto de sentencias para la elaboración de un modelo *k-NN* con k

Los resultados que se obtuvieron son los siguientes:

CCR	CCR 10-fold
0.7600473	0.6483362

Cuadro 70: Métricas de rendimiento evaluadas para el modelo 7-NN elaborado con todas las variables del *dataset*

Como primer modelo obtenido, el rendimiento obtenido es apreciable, no obstante, todavía se pretende mejorar este valor de *Accuracy*.

A continuación, se pretende comprobar los efectos de un escalado sobre los resultados, dando lugar al modelo *knn.model.fit2*:

```

1 # Modelo con k=7, con atributos escalados
2 vehicle.rescaled <- vehicle
3 vehicle.rescaled[,-ncol(vehicle)] <- scale(vehicle.
   rescaled[,-ncol(vehicle)])
4
5 knn.model.fit2 <- knn(train=vehicle.rescaled[, -ncol(
  vehicle.rescaled)],
   test=vehicle.rescaled[, -ncol(
  vehicle.rescaled)],
   cl=vehicle.rescaled[, ncol(vehicle.
  rescaled)], k=7)
6
7 # Evaluar error CCR train
8 cat('Accuracy sobre el conjunto de train: ', postResample
  (knn.model.fit2,
   vehicle.
   rescaled$Class)[1],
   fill=T)
9 cat('Accuracy sobre 10-fold: ', mean(sapply(1:10, run_knn_
  k_fold_vehicle,
   fold_basename
  , 7,
   vehicle)[1:ncol(vehicle)-1])),
   colnames(
  vehicle)[1:ncol(vehicle)-1]),
   fill=T)
10
11
12
13
14
15
16

```

Script 32: Conjunto de sentencias para la elaboración de un modelo k-NN con $k=7$, aplicando escalado sobre el dataset

Los resultados se muestran a continuación:

CCR	CCR 10-fold
0.8144208	0.7080895

Cuadro 71: Métricas de rendimiento evaluadas para el modelo 7-NN elaborado con todas las variables del *dataset* a las cuales se ha aplicado un escalado

Se puede apreciar que, tras aplicar el escalado de todas las variables independientes del *dataset*, el *accuracy* de entrenamiento y el *accuracy* medido sobre *10-fold* han mejorado notablemente, con lo que mejora la capacidad de generalización del modelo.

Finalmente, se realiza una batería de experimentos generando diferentes modelos con diferentes valores de k y, evaluando el rendimiento obtenido para cada uno de ellos. Los valores de k que se consideran son los siguientes:

1,3,7,13,21,51,75 y 103.

k	CCR	CCR 10-fold
1	1	0.7071429
3	0.8427896	0.6857143
7	0.8108747	0.7261905
13	0.7695035	0.7119048
21	0.7411348	0.7142857
51	0.6808511	0.6523810
75	0.6583924	0.6214286
103	0.6146572	0.6047619

Cuadro 72: Métricas de rendimiento evaluadas para los diferentes modelos generados para cada valor de k , a partir del modelo que reescalaba todas las variables independientes. Se ha resaltado en negro los valores óptimos.

Tras la ejecución de la batería de pruebas, se aprecia que el modelo que ha permitido obtener mejor rendimiento, es aquel que considera un valor de $k=7$ y aplica escalado de todas las variables independientes.

3.3.2. Elaboración de modelos basados en LDA

A continuación, se estudiarán y elaborará modelos basados en *LDA* (*Linear Discriminant Analysis*).

Los modelos *LDA* se formulan en base a una serie de asunciones las cuales es necesario garantizar, para asegurar que el modelo rendirá correctamente:

Una de las primeras asunciones es la normalidad en las distribuciones de cada clase. Como ya se analizó en la sección de análisis exploratorio, las distribuciones de las variables diferían de la distribución Normal, por lo que sabiendo que las distribuciones de las variables se hallan incluídas en ellas, no parece haber indicios de que las distribuciones de cada clase tiendan a una Normal.

Otra de las asunciones de los modelos *LDA*, es la equidad en las varianzas de las distribuciones. Para analizar si las distribuciones de las clases presentan varianzas similares, se recoge en la siguiente tabla 75, las varianzas calculadas para cada variable en cada clase:

Variable	bus	opel	saab	ban
Compactness	74.27035	67.74504	82.47154	14.99477
Circularity	25.30861	52.34899	46.49996	16.67179
Distance_circularity	146.2301	242.9445	289.1749	118.1319
Radius_ratio	934.7788	983.1312	948.9374	893.6824
Praxis_aspect_ratio	77.49064	24.56309	18.67857	129.51723
Max_length_aspect_ratio	22.631146	3.923008	4.628094	52.195320
Scatter_ratio	1112.6769	1077.2759	993.7505	195.8761
Elongatedness	42.26789	59.66261	56.05120	21.79433
Praxis_rectangular	7.443897	6.517996	6.016854	1.063347
Length_rectangular	110.0904	329.7163	260.3255	121.3376
Major_variance	1155.7666	831.8012	779.5277	387.3029
Minor_variance	37302.408	29742.193	26619.959	3124.212
Gyration_radius	976.3984	1205.5788	1140.8233	520.8071
Major_skewness	59.05832	25.80926	28.00026	78.62763
Minor_skewness	10.37188	26.97022	33.83687	21.76961
Minor_kurtosis	47.23638	103.35999	101.34042	38.97934
Major_kurtosis	53.60041	31.28237	24.91799	40.70342
Hollows_ratio	62.69989	34.17670	43.69735	53.95340

Cuadro 73: Varianzas de todas las variables independientes del *dataset* calculadas por clase.

Comparando las varianzas en cada fila, podemos comprobar de forma clara que las distribuciones de las clases presentan diferentes varianzas.

No obstante, el modelo permite salvar esta asunción si se da que las distribuciones de las clases presentan equidad en las covarianzas. Para comprobar si existe equidad en las covarianzas de las distribuciones, hacemos uso del test de *BoxM*:

```

1 # Analizar las equidad en las covarianzas con el test de
  BoxM
2 test.boxm <- BoxM(vehicle[,-ncol(vehicle)], group=
  vehicle$Class)
3 cat('p-value asociado al test BoxM: ',test.boxm$p.value)
```

Script 33: Conjunto de sentencias para la realización de un test BoxM sobre las distribuciones de clase

p-value asociado al test BoxM: 0

El *p-value* asociado al test no es significativo y, por lo tanto, nos lleva a rechazar la hipótesis nula y a negar que las covarianzas sean homogéneas en

las distribuciones de clase, en otras palabras, tampoco se cumple la asunción de que las covarianzas de las distribuciones de las clases son iguales.

Llegados a este punto, realizaremos modelos basados en *LDA* sin poder garantizar el rendimiento del modelo.

En el análisis exploratorio que se realizó en el *dataset*, se pudo comprobar que las distribuciones de las variables del *dataset* no tienden a una distribución Normal, por lo que no se da esta condición ideal, no obstante, ello no invalida la construcción de modelos basados en *LDA*.

Primeramente elaboramos un modelo excluyendo únicamente todas las variables que guarden una correlación lineal de al menos 0.9 con cualquier otra variable del *dataset*, dando lugar al modelo *lda.model.fit1*, en el cual las siguientes variables han sido excluidas: *Scatter_ratio*, *Praxis_rectangular*, *Length_rectangular*, *Minor_ariance* y *Gyration_radius*).

```

1 # Class ~ .
2 lda.model.fit1 <- lda(Class~.-Scatter_ratio-Praxis_
3                         rectangular-Length_rectangular-
4                         Minor_variance-Gyration_radius ,
5                         data=vehicle)
6 lda.model.fit1
7
8 lda.model.fit1.pred <- predict(lda.model.fit1, vehicle)
9 cat('CCR medido sobre entrenamiento:', postResample(lda.
10     model.fit1.pred$class ,
11
12     vehicle$Class)[1], fill=T)
13 cat('Accuracy sobre 10-fold: ', mean(sapply(1:10, run_lda_
14     k_fold_vehicle,
15                         fold_basename
16     , lda.model.fit1)), fill=T)
```

Script 34: Conjunto de sentencias para la elaboración del modelo *lda.model.fit1*

El rendimiento del modelo se expone a continuación:

Modelo LDA	CCR	CCR 10-fold
lda.model.fit1	0.7765957	0.7588655

Cuadro 74: Modelos LDA ignorando los atributos *Scatter_ratio*, *Praxis_rectangular*, *Length_rectangular*, *Minor_ariance* y *Gyration_radius*.

Por último, se elaborará un último modelo con todas las variables independientes del *dataset* (modelo *lda.model.fit2*)

```

1 # Class ~ .
2 lda.model.fit2 <- lda(Class~, data=vehicle)
3 lda.model.fit2
4
5 lda.model.fit2.pred <- predict(lda.model.fit2, vehicle)
6 cat('CCR medido sobre entrenamiento:', postResample(lda.
7     model.fit2.pred$class,
8     vehicle$Class)[1], fill=T)
9 cat('Accuracy sobre 10-fold: ', mean(sapply(1:10, run_lda_
10     k_fold_vehicle,
11                     fold_basename
12     , lda.model.fit2)), fill=T)

```

Script 35: Conjunto de sentencias para la elaboración del modelo *lda.model.fit2*

Con este modelo se obtuvo las siguientes métricas de precisión:

Modelo LDA	CCR	CCR 10-fold
lda.model.fit2	0.7978723	0.7813305

Cuadro 75: Modelos LDA desarrollados con todas las variables del *dataset*.

Se aprecia claramente que, este último modelo ofrece un mejor rendimiento.

3.3.3. Elaboración de modelos basados en QDA

Por último, se estudiarán y elaborarán modelos basados en *QDA* (*Quadratic Discriminant Analysis*).

Los modelos basados en *QDA* no requieren asumir que las varianzas y/o covarianzas de las distribuciones de clase sean similares, lo cual para este problema resulta idóneo.

No obstante, una de las asunciones de este modelo es la ausencia de niveles de correlación entre las variables para cada clase. En la siguiente figura se muestran las correlaciones entre los pares de variables para cada clase

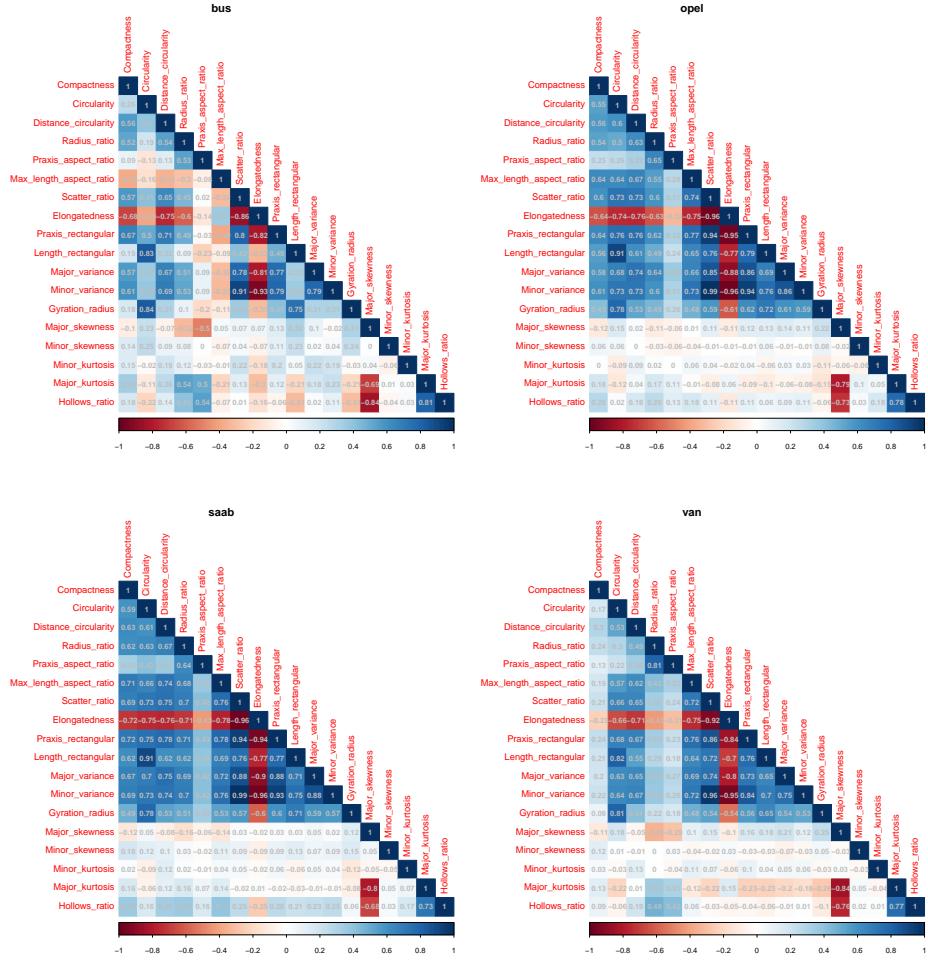


Figura 76: Cuadro con correlograma que analiza la correlación lineal entre las variables haciendo una separación por clase, como medida de correlación se ha usado el coeficiente de *Kendall*.

Se observa que existen fuertes correlaciones en muchas de las variables, por lo que esta sunción no se da y, por tanto, esto nos lleva a dudar del rendimiento del modelo.

En cualquier caso, se elabora un modelo con todas las variables de *dataset*:

Modelo LDA	CCR	CCR 10-fold
qda.model.fit1	0.9160757	0.8522409

Cuadro 76: Modelos QDA desarrollados con todas las variables del *dataset*.

Se observa que el rendimiento del modelo es elevado.

3.3.4. Comparación de modelos

Finalmente, procedemos a comparar los mejores modelos obtenidos en este epígrafe, esta comparación se efectuará en función de las métricas de precisión que se han usado hasta ahora: *CCR* medido sobre el conjunto de *train* y el *CCR* medido sobre *10-fold*.

Descripción del modelo	CCR train	CCR 10-fold
Modelo 7-NN que aplica reescalamiento sobre las variables	0.8108747	0.7261905
Modelo LDA con todas las variables del <i>dataset</i> como variables dependientes	0.7978723	0.7813305
Modelo QDA con todas las variables del <i>dataset</i> como variables dependientes	0.9160757	0.8522409

Cuadro 77: Comparación los mejores modelos obtenidos

De entre todos los modelos generados, el modelo basado en *QDA* es el que obtiene mejores resultados con diferencia, seguido del modelo *LDA* y, finalmente por el modelo basado en *k-NN*.

En este caso, aunque no se cumplan todas las asunciones del modelo *QDA*, que este modelo sea flexible ante varianzas diferentes y distribuciones no normales para las clases, le permite obtener un mejor rendimiento frente a *LDA* que, al no cumplirse estas asunciones en el *dataset*, no obtiene un buen rendimiento.

Por último, aunque el modelo basado en *k-NN* ofrezca un mejor rendimiento sobre el conjunto de *train* que *LDA*, se aprecia que este último ofrece mayor capacidad de generalización.

A. Script1.R

Este script fue usado para realizar el análisis exploratorio de datos sobre el *dataset house*:

```
1 # #####  
2 # Nombre del script: script1.R  
3 # Desarrollado por: Nicolás Cubero Torres  
4 # Descripción: Script para el análisis exploratorio del  
# dataset house  
5 # Nota: Este script ha sido desarrollado para ejecutarse  
# de forma interactiva  
6 # línea por línea  
7 # #####  
8  
9 # Librerías cargadas  
10 library('moments')  
11 library('ggplot2')  
12 library('GGally')  
13 library('dplyr')  
14 library('corrplot')  
15  
16 # Función para cargar house  
17 read.house.dataset <- function(filename) {  
18     # Cargar datos  
19     dat <- read.table(filename, comment.char="@", sep=',')  
20  
21     # Asignar nombres  
22     names(dat) <- c('P1', 'P5p1', 'P6p2', 'P11p4', 'P14p9'  
23     , 'P15p1', 'P15p3',  
24         'P16p2', 'P18p2', 'P27p4', 'H2p2',  
25         'H8p2', 'H10p1', 'H13p1',  
26         'H18pA', 'H40p4', 'Price')  
27  
28     return(dat)  
29 }  
30  
31 # Cargar el dataset house  
32 house <- read.house.dataset('./Datasets\ Regresion/house  
# /house.dat')  
33  
34 # Obtener información sobre su estructura  
35 str(house)  
36 head(house)
```

```

36 # Obtener información sobre Missing Values
37 any(is.na(house))
38
39 # Determinar los estadísticos de posición:
40 # Valores mínimo y máximos, media, 1er cuartil, mediana,
41 # 3er cuartil
41 cat('Estadísticos de posición: Valores mínimo y máximos,
42     media, 1er cuartil ,',
43     ' mediana, 3er cuartil', fill=T)
43 summary(house)
44
45 # Determinar los estadísticos de dispersión: desviación
45 # típica
46 cat('Desviación típica de los atributos', fill=T)
47 apply(house, MARGIN=2, FUN=sd)
48
49 # Determinar coeficientes de Skew y Kurtosis
50 cat('Coeficientes de Skew y Kurtosis', fill=T)
51 skew_kurtosis <- apply(house, MARGIN=2, FUN=function(x)
51   {c(skewness(x),
52
52   kurtosis(x))})
53 rownames(skew_kurtosis) <- c('Skew', 'kurtosis')
54 skew_kurtosis
55
56 # Variable P1
57
58 # Diagrama boxplot
59 graf <- ggplot(house, aes(x='P1', y=P1)) +
60   geom_boxplot(outlier.alpha=0.4, outlier.colour='
60   red', outlier.shape=8) +
61   labs(cation='center', y=' ', x=' ')
62 graf
63
64 # Diagrama boxplot en escala logarítmica
65 hraf <- ggplot(house, aes(x='P1', y=P1)) +
66   geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
66   outlier.shape=8) +
67   scale_y_log10() +
68   labs(cation='center', y=' ', x=' ')
69 hraf
70
71 # Aislar la distribución comprendida entre [Q1-1.5*IQR,
71 # Q3+1.5*IQR]
72 quantiles.P1 <- quantile(house$P1)
73 iqr.P1 <- IQR(house$P1)
74
75 centro.house.P1 <- house %>% select(P1) %>%

```

```

76 filter(P1 > (quantiles.P1[2]-1.5*iqr.P1) & (P1 <
77   quantiles.P1[4]*1.5))
78
79 # Obtener estadísticos de posición y número de medidas
80 cat('Rango [Q1-1.5*IQR, Q3+1.5*IQR]:', fill=T)
81 centro.house.P1 %>% summarise(min=min(centro.house.P1$P1
82   ),
83     q1=quantile(centro.
84     house.P1$P1)[2],
85     median=median(centro.
86     house.P1$P1),
87     mean=mean(centro.
88     house.P1$P1)[4],
89     q3=quantile(centro.
90     house.P1$P1),
91     max=max(centro.house
92     .P1$P1),
93     count=n(),
94     ratio=n()/length(
95     house$P1))
96
97 # Obtener los outliers inferiores a Q1-1.5*IQR
98 cat('Rango (-Inf, Q1-1.5*IQR):', fill=T)
99 low.outliers.house.P1 <- house %>% select(P1) %>%
100   filter(P1 < (quantiles.P1[2]-1.5*iqr.P1))
101
102 # Obtener estadísticos de posición y número de medidas
103 low.outliers.house.P1 %>% summarise(min=min(low.outliers
104   .house.P1),
105     q1=quantile(low.
106     outliers.house.P1)[2],
107     median=median(low.
108     outliers.house.P1),
109     mean=mean(low.
110     outliers.house.P1)[4],
111     q3=quantile(low.
112     outliers.house.P1),
113     max=max(low.outliers
114     .house.P1),
115     count=n(),
116     ratio=n()/length(
117     house$P1))
118
119 # Obtener los outliers superiores a Q3+1.5*IQR
120 cat('Rango (Q3+1.5*IQR, +Inf):', fill=T)
121 upper.outliers.house.P1 <- house %>% select(P1) %>%
122   filter(P1 > (quantiles.P1[4]+1.5*iqr.P1))
123
124 # Obtener estadísticos de posición y número de medidas

```

```

110 upper.outliers.house.P1 %>% summarise(min=min(upper.
111   outliers.house.P1$P1),
112     q1=quantile(upper.outliers
113       .house.P1$P1)[2],
114       median=median(upper.
115         outliers.house.P1$P1),
116       mean=mean(upper.outliers.
117         house.P1$P1),
118       q3=quantile(upper.outliers
119         .house.P1$P1)[4],
120       max=max(upper.outliers.
121         house.P1$P1),
122       count=n(),
123       ratio=n()/length(house$P1)
124     )
125
126 # Histograma del centro de distribución
127 graf.high_density.P1 <- ggplot(centro.house.P1, aes(P1))
128   +
129   geom_histogram(binwidth=15, color='red', alpha=0.4) +
130   labs(y='frecuencia absoluta')
131 graf.high_density.P1
132
133 # Histograma del resto de la distribución
134 graf.upper_outliers.P1 <- ggplot(upper.outliers.house.P1
135   , aes(P1)) +
136   geom_histogram(binwidth=1000, colour='red') +
137   labs(y='frecuencia absoluta')
138 graf.upper_outliers.P1
139
140 # Gráfica P5p1
141
142 # Diagrama boxplot
143 graf <- ggplot(house, aes(x='P5p1', y=P5p1)) +
144   geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
145     outlier.shape=8) +
146   labs(cation='center', y=' ', x=' ')
147 graf
148
149 # Histograma
150 graf <- ggplot(house, aes(P5p1)) +
151   geom_histogram(binwidth=0.0025, color='gray58') +
152   labs(y='frecuencia absoluta')
153 graf
154
155 # Gráfica P6p1
156
157 # Diagrama boxplot
158 graf <- ggplot(house, aes(x='P6p1', y=P6p2)) +

```

```

149  geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
150    outlier.shape=1) +
151  labs(caption='center', y=' ', x=' ')
151 graf
152
153 # Histograma
154 graf <- ggplot(house, aes(P6p2)) +
155   geom_histogram(binwidth=0.0025, color='gray58')
156   labs(y='frecuencia absoluta')
157 graf
158
159 # Gráfica P11p4
160
161 # Diagrama boxplot
162 graf <- ggplot(house, aes(x='P11p4', y=P11p4)) +
163   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
164     outlier.shape=1) +
165   labs(caption='center', y=' ', x=' ')
165 graf
166
167 # Histograma
168 graf <- ggplot(house, aes(P11p4)) +
169   geom_histogram(binwidth=0.0025, color='gray58')
170   labs(y='frecuencia absoluta')
171 graf
172
173 # Gráfica P14p9
174
175 # Diagrama boxplot
176 graf <- ggplot(house, aes(x='P14p9', y=P14p9)) +
177   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
178     outlier.shape=1) +
179   labs(caption='center', y=' ', x=' ')
179 graf
180
181 # Histograma
182 graf <- ggplot(house, aes(P14p9)) +
183   geom_histogram(binwidth=0.00125, color='gray58')
184   labs(y='frecuencia absoluta')
185 graf
186
187 # Gráfica P15p1
188
189 # Diagrama boxplot
190 graf <- ggplot(house, aes(x='P15p1', y=P15p1)) +
191   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
192     outlier.shape=1) +
193   labs(caption='center', y=' ', x=' ')
193 graf

```

```

194
195 # Histograma
196 graf <- ggplot(house, aes(P15p1)) +
197   geom_histogram(binwidth=0.00064, color='gray58')
198 labs(y='frecuencia absoluta')
199 graf
200
201 # Gráfica P15p3
202
203 # Diagrama boxplot
204 graf <- ggplot(house, aes(x='P15p3', y=P15p3)) +
205   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
206     outlier.shape=1) +
207   labs(cation='center', y='', x='')
208 graf
209
210 # Histograma
211 graf <- ggplot(house, aes(P15p3)) +
212   geom_histogram(binwidth=0.00064, color='gray58')
213 labs(y='frecuencia absoluta')
214 graf
215
216 # Gráfica P16p2
217
218 # Diagrama boxplot
219 graf <- ggplot(house, aes(x='P16p2', y=P16p2)) +
220   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
221     outlier.shape=1) +
222   labs(cation='center', y='', x='')
223 graf
224
225 # Histograma
226 graf <- ggplot(house, aes(P16p2)) +
227   geom_histogram(binwidth=0.00125, color='gray58')
228 labs(y='frecuencia absoluta')
229 graf
230
231 # P18p2
232
233 # Diagrama boxplot
234 graf <- ggplot(house, aes(x='P18p2', y=P18p2)) +
235   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
236     outlier.shape=1) +
237   labs(cation='center', y='', x='')
238 graf
239
240 # Histograma
241 graf <- ggplot(house, aes(P18p2)) +
242   geom_histogram(binwidth=0.000125, color='gray58') +

```

```

240 labs(y='frecuencia absoluta')
241 graf
242
243 # Histograma del centro de la distribución de P18p2
244 iqr.P18p2 <- IQR(house$P18p2)
245 quantiles.P18p2 <- quantile(house$P18p2)
246
247 graf <- ggplot(house, aes(P18p2)) +
248   geom_histogram(binwidth=0.0001, color='gray58') +
249   labs(y='frecuencia absoluta') +
250   xlim(quantiles.P18p2[2]-1.5*iqr.P18p2, quantiles.P18p2
251     [4]+1.5*iqr.P18p2)
252 graf
253
254 # P27p4
255
256 # Diagrama boxplot
257 graf <- ggplot(house, aes(x='P27p4', y=P27p4)) +
258   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
259     outlier.shape=1) +
260   labs(cation='center', y='', x='')
261 graf
262
263 # Histograma
264 graf <- ggplot(house, aes(P27p4)) +
265   geom_histogram(binwidth=0.000125, color='gray58') +
266   labs(y='frecuencia absoluta')
267 graf
268
269 # H2p2
270
271 # Diagrama boxplot
272 graf <- ggplot(house, aes(x='H2p2', y=H2p2)) +
273   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
274     outlier.shape=1) +
275   labs(cation='center', y='', x='')
276 graf
277
278 # Histograma
279 graf <- ggplot(house, aes(H2p2)) +
280   geom_histogram(binwidth=0.0005, color='gray58') +
281   labs(y='frecuencia absoluta')
282 graf
283
284 # H8p2
285
286 # Diagrama boxplot
287 graf <- ggplot(house, aes(x='H8p2', y=H8p2)) +

```

```

285 geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
286   outlier.shape=1) +
287   labs(caption='center', y=' ', x=' ')
288 graf
289
290 # Histograma
291 graf <- ggplot(house, aes(H8p2)) +
292   geom_histogram(binwidth=0.0005, color='gray58') +
293   labs(y='frecuencia absoluta')
294 graf
295
296 # H10p1
297
298 # Diagrama boxplot
299 graf <- ggplot(house, aes(x='H10p1', y=H10p1)) +
300   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
301     outlier.shape=1) +
302   labs(caption='center', y=' ', x=' ')
303 graf
304
305 # Histograma
306 graf <- ggplot(house, aes(H10p1)) +
307   geom_histogram(binwidth=0.0005, color='gray58') +
308   labs(y='frecuencia absoluta')
309 graf
310
311 # H13p1
312
313 # Diagrama boxplot
314 graf <- ggplot(house, aes(x='H13p1', y=H13p1)) +
315   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
316     outlier.shape=1) +
317   labs(caption='center', y=' ', x=' ')
318 graf
319
320 # Histograma
321 graf <- ggplot(house, aes(H13p1)) +
322   geom_histogram(binwidth=0.0005, color='gray58') +
323   labs(y='frecuencia absoluta')
324 graf
325
326 # H18pA
327
328 # Diagrama boxplot
329 graf <- ggplot(house, aes(x='H18pA', y=H18pA)) +
330   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
331     outlier.shape=1) +
332   labs(caption='center', y=' ', x=' ')
333 graf

```

```

330
331 # Histograma
332 graf <- ggplot(house, aes(H18pA)) +
333   geom_histogram(binwidth=0.0025, color='gray58') +
334   labs(y='frecuencia absoluta')
335 graf
336
337 # H40p4
338
339 # Diagrama boxplot
340 graf <- ggplot(house, aes(x='H40p4', y=H40p4)) +
341   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
342     outlier.shape=1) +
343   labs(cation='center', y=' ', x=' ')
344 graf
345
346 # Histograma
347 graf <- ggplot(house, aes(H40p4)) +
348   geom_histogram(binwidth=0.0025, color='gray58') +
349   labs(y='frecuencia absoluta')
350 graf
351
352 # Price
353
354 # Diagrama boxplot
355 graf <- ggplot(house, aes(x='Price', y=Price)) +
356   geom_boxplot(outlier.alpha=0.1, outlier.colour='red',
357     outlier.shape=1) +
358   labs(cation='center', y=' ', x=' ')
359 graf
360
361 # Histograma
362 graf <- ggplot(house, aes(Price)) +
363   geom_histogram(binwidth=200, color='gray58') +
364   labs(y='frecuencia absoluta')
365 graf
366
367 # Estudiar los pares de relaciones
368 png('relaciones_pares_variables.png', width=480*5,
      height = 240*5, res=120)
369 ggpairs(house, mapping=ggplot2::aes(colour='red', alpha
      =0.05),
      upper = list(continuous = "cor", corMethod = "
      kendall"))
370 dev.off()
371
372 # Estudiar la relación entre las variables P1 y P5p1
373 graf.p1.p5p1 <- ggplot(house, aes(x=P1, y=P5p1)) +
      geom_point(color='red', alpha=0.3)

```

```

374 graf.p1.p5p1
375
376 # Dibujar un gráfico de correlaciones basados en
377 # Spearman
377 corr <- cor(house, method='kendall')
378
379 pdf('Correlograma_kendall_house.pdf', width=14, height =
380     7)
380 par(mfrow=c(1,2))
381 corrplot(corr, method='ellipse', type='lower')
382 corrplot(corr, method='number', type='lower', number.cex
383     = 0.80)
383 dev.off()

```

B. Script2.R

Este *script* fue usado para la generación de todos los modelos de regresión para el *dataset house*

```

1 #
# ##########
2 # Nombre del script: script2.R
3 # Desarrollado por: Nicolás Cubero Torres
4 # Descripción: Script para el desarrollo de modelos de
#   regresión sobre el
5 # dataset house
6 # Nota: Este script ha sido desarrollado para ejecutarse
#       de forma interactiva
7 #       línea por línea
8 #
# ##########
9 # Librerías cargadas
10 library('ggplot2')
11 library('gridExtra')
12
13 library('kknn')
14
15 read.house.dataset <- function(filename) {
16   # Cargar datos
17   dat <- read.table(filename, comment.char="@", sep=',')
18
19   # Asignar nombres
20   names(dat) <- c('P1', 'P5p1', 'P6p2', 'P11p4', 'P14p9',
21     'P15p1', 'P15p3',
22     'P16p2', 'P18p2', 'P27p4', 'H2p2', 'H8p2', 'H10p1', 'H13p1',

```

```

22             'H18pA', 'H40p4', 'Price')
23
24     return(dat)
25 }
26
27 # Cargar el dataset house
28 house <- read.house.dataset('./Datasets\ Regresion/house
29     /house.dat')
30
31 # Definición de la función para evaluar un modelo lineal
32 evaluate.rmse.house <- function(model) {
33     yprime = predict(model, house)
34     sqrt(sum(abs(house$Price-yprime)^2)/length(yprime)) # 
35         Calcular RMSE
36 }
37
38 # Modelos lineales
39 lineal.simple.P1 <- lm(Price ~ P1, data=house)
40 lineal.simple.P5p1 <- lm(Price ~ P5p1, data=house)
41 lineal.simple.P6p2 <- lm(Price ~ P6p2, data=house)
42 lineal.simple.P11p4 <- lm(Price ~ P11p4, data=house)
43 lineal.simple.P14p9 <- lm(Price ~ P14p9, data=house)
44 lineal.simple.P14p9 <- lm(Price ~ P14p9, data=house)
45 lineal.simple.P15p1 <- lm(Price ~ P15p1, data=house)
46 lineal.simple.P15p3 <- lm(Price ~ P15p3, data=house)
47 lineal.simple.P16p2 <- lm(Price ~ P16p2, data=house)
48 lineal.simple.P18p2 <- lm(Price ~ P18p2, data=house)
49 lineal.simple.P27p4 <- lm(Price ~ P27p4, data=house)
50 lineal.simple.H2p2 <- lm(Price ~ H2p2, data=house)
51 lineal.simple.H8p2 <- lm(Price ~ H8p2, data=house)
52 lineal.simple.H10p1 <- lm(Price ~ H10p1, data=house)
53 lineal.simple.H13p1 <- lm(Price ~ H13p1, data=house)
54 lineal.simple.H18pA <- lm(Price ~ H18pA, data=house)
55 lineal.simple.H40p4 <- lm(Price ~ H40p4, data=house)
56
57 # Obtener información sobre los modelos
58 summary(lineal.simple.P1)
59 evaluate.rmse.house(lineal.simple.P1)
60
61 summary(lineal.simple.P5p1)
62 evaluate.rmse.house(lineal.simple.P5p1)
63
64 summary(lineal.simple.P6p2)
65 evaluate.rmse.house(lineal.simple.P6p2)
66
67 summary(lineal.simple.P11p4)
68 evaluate.rmse.house(lineal.simple.P11p4)
69
70 summary(lineal.simple.P14p9)

```

```

69 evaluate.rmse.house(lineal.simple.P14p9)
70
71 summary(lineal.simple.P15p1)
72 evaluate.rmse.house(lineal.simple.P15p1)
73
74 summary(lineal.simple.P15p3)
75 evaluate.rmse.house(lineal.simple.P15p3)
76
77 summary(lineal.simple.P16p2)
78 evaluate.rmse.house(lineal.simple.P16p2)
79
80 summary(lineal.simple.P18p2)
81 evaluate.rmse.house(lineal.simple.P18p2)
82
83 summary(lineal.simple.P27p4)
84 evaluate.rmse.house(lineal.simple.P27p4)
85
86 summary(lineal.simple.H2p2)
87 evaluate.rmse.house(lineal.simple.H2p2)
88
89 summary(lineal.simple.H8p2)
90 evaluate.rmse.house(lineal.simple.H8p2)
91
92 summary(lineal.simple.H10p1)
93 evaluate.rmse.house(lineal.simple.H10p1)
94
95 summary(lineal.simple.H13p1)
96 evaluate.rmse.house(lineal.simple.H13p1)
97
98 summary(lineal.simple.H18pA)
99 evaluate.rmse.house(lineal.simple.H18pA)
100
101 summary(lineal.simple.H40p4)
102 evaluate.rmse.house(lineal.simple.H40p4)
103
104 # Representación de los modelos elaborados
105 graf.simple.model.P1 <- ggplot(house, aes(y=Price, x=P1))
106   +
107   geom_point(colour='red', alpha=0.2) +
108   geom_abline(intercept=lineal.simple.P1$coefficients
109               [2],
110               slope=lineal.simple.P1$coefficients[1],
111               colour='deepskyblue',
112               size=0.9)
113 graf.simple.model.P1
114
115 graf.simple.model.P5p1 <- ggplot(house, aes(y=Price, x=
116                                         P5p1)) +
117   geom_point(colour='red', alpha=0.2) +

```

```

115 geom_abline(intercept=lineal.simple.P5p1$coefficients
116   [2] ,
117     slope=lineal.simple.P5p1$coefficients[1] ,
118     colour='deepskyblue')
119 graf.simple.model.P5p1
120
121 graf.simple.model.P6p2 <- ggplot(house , aes(y=Price , x=
122   P6p2)) +
123   geom_point(colour='red' , alpha=0.2) +
124   geom_abline(intercept=lineal.simple.P6p2$coefficients
125   [2] ,
126     slope=lineal.simple.P6p2$coefficients[1] ,
127     colour='deepskyblue')
128 graf.simple.model.P6p2
129
130 graf.simple.model.P11p4 <- ggplot(house , aes(y=Price , x=
131   P11p4)) +
132   geom_point(colour='red' , alpha=0.2) +
133   geom_abline(intercept=lineal.simple.P11p4$coefficients
134   [2] ,
135     slope=lineal.simple.P11p4$coefficients[1] ,
136     colour='deepskyblue')
137 graf.simple.model.P11p4
138
139 graf.simple.model.P14p9 <- ggplot(house , aes(y=Price , x=
140   P14p9)) +
141   geom_point(colour='red' , alpha=0.2) +
142   geom_abline(intercept=lineal.simple.P14p9$coefficients
143   [2] ,
144     slope=lineal.simple.P14p9$coefficients[1] ,
145     colour='deepskyblue')
146 graf.simple.model.P14p9
147
148 graf.simple.model.P15p1 <- ggplot(house , aes(y=Price , x=
149   P15p1)) +
150   geom_point(colour='red' , alpha=0.2) +
151   geom_abline(intercept=lineal.simple.P15p1$coefficients
152   [2] ,
153     slope=lineal.simple.P15p1$coefficients[1] ,
154     colour='deepskyblue')

```

```

153 graf.simple.model.P15p3
154
155 graf.simple.model.P16p2 <- ggplot(house, aes(y=Price, x=
156   P16p2)) +
157   geom_point(colour='red', alpha=0.2) +
158   geom_abline(intercept=lineal.simple.P16p2$coefficients
159     [2],
160       slope=lineal.simple.P16p2$coefficients[1],
161       colour='deepskyblue')
160 graf.simple.model.P16p2
161
162 graf.simple.model.P18p2 <- ggplot(house, aes(y=Price, x=
163   P18p2)) +
164   geom_point(colour='red', alpha=0.2) +
165   geom_abline(intercept=lineal.simple.P18p2$coefficients
166     [2],
167       slope=lineal.simple.P18p2$coefficients[1],
168       colour='deepskyblue')
167 graf.simple.model.P18p2
168
169 graf.simple.model.P27p4 <- ggplot(house, aes(y=Price, x=
170   P27p4)) +
171   geom_point(colour='red', alpha=0.2) +
172   geom_abline(intercept=lineal.simple.P27p4$coefficients
173     [2],
174       slope=lineal.simple.P27p4$coefficients[1],
175       colour='deepskyblue')
174 graf.simple.model.P27p4
175
176 graf.simple.model.H2p2 <- ggplot(house, aes(y=Price, x=
177   H2p2)) +
178   geom_point(colour='red', alpha=0.2) +
179   geom_abline(intercept=lineal.simple.H2p2$coefficients
180     [2],
181       slope=lineal.simple.H2p2$coefficients[1],
182       colour='deepskyblue')
181 graf.simple.model.H2p2
182
183 graf.simple.model.H8p2 <- ggplot(house, aes(y=Price, x=
184   H8p2)) +
185   geom_point(colour='red', alpha=0.2) +
186   geom_abline(intercept=lineal.simple.H8p2$coefficients
187     [2],
188       slope=lineal.simple.H8p2$coefficients[1],
189       colour='deepskyblue')
188 graf.simple.model.H8p2
189
190 graf.simple.model.H10p1 <- ggplot(house, aes(y=Price, x=
191   H10p1)) +

```

```

191 geom_point(colour='red', alpha=0.2) +
192 geom_abline(intercept=lineal.simple.H10p1$coefficients
193 [2],
194 slope=lineal.simple.H10p1$coefficients[1],
195 colour='deepskyblue')
196 graf.simple.model.H10p1
197
198 graf.simple.model.H13p1 <- ggplot(house, aes(y=Price, x=
199 H13p1)) +
200 geom_point(colour='red', alpha=0.2) +
201 geom_abline(intercept=lineal.simple.H13p1$coefficients
202 [2],
203 slope=lineal.simple.H13p1$coefficients[1],
204 colour='deepskyblue')
205 graf.simple.model.H13p1
206
207 graf.simple.model.H18pA <- ggplot(house, aes(y=Price, x=
208 H18pA)) +
209 geom_point(colour='red', alpha=0.2) +
210 geom_abline(intercept=lineal.simple.H18pA$coefficients
211 [2],
212 slope=lineal.simple.H18pA$coefficients[1],
213 colour='deepskyblue')
214 graf.simple.model.H18pA
215
216 graf.simple.model.H40p4 <- ggplot(house, aes(y=Price, x=
217 H18pA)) +
218 geom_point(colour='red', alpha=0.2) +
219 geom_abline(intercept=lineal.simple.H40p4$coefficients
220 [2],
221 slope=lineal.simple.H40p4$coefficients[1],
222 colour='deepskyblue')
223 graf.simple.model.H40p4
224
225 # Representar todas las gráficas
226 png('modelos_lineales.png', width=240*5, height = 480*5,
227 res=120)
228 grid.arrange(graf.simple.model.P1, graf.simple.model.
229 P5p1,
230 graf.simple.model.P6p2, graf.simple.model.
231 P11p4,
232 graf.simple.model.P14p9, graf.simple.model.
233 P15p1,
234 graf.simple.model.P15p3, graf.simple.model.
235 P16p2,
236 graf.simple.model.P18p2, graf.simple.model.
237 P27p4,
238 graf.simple.model.H2p2, graf.simple.model.
239 H8p2,
```

```

226           graf.simple.model.H10p1, graf.simple.model.
227           H13p1,
228           graf.simple.model.H18pA, graf.simple.model.
229           H40p4, nrow=8)
230 dev.off()
231
230 # Los 5 mejores modelos encontrados:
231 # 1º Price ~ P27p4
232 # 2º Price ~ P11p4
233 # 3º Price ~ H13p1
234 # 4º & Price ~ H40p4
235 # 5º & Price ~ P16p2
236
237 # Evaluar los modelos mediante validación cruzada 5-fold
238 run_k_fold <- function(i, x, model, type='lineal', tt =
239   "test") {
240   # Cargar conjuntos de entrenamiento
241   file <- paste(x, "-5-", i, "tra.dat", sep=" ");
242   x_tra <- read.house.dataset(file)
243
244   # Cargar conjuntos de test
245   file <- paste(x, "-5-", i, "tst.dat", sep=" ");
246   x_tst <- read.house.dataset(file)
247
248   if (tt == "train") { test <- x_tra }
249   else { test <- x_tst }
250
251   # Entrenar el modelo sobre el conjunto de train
252   form <- terms(model)
253   model.eval <- lm(formula=form, data=x_tra)
254
255   # Evaluación del RMSE sobre test
256   yprime=predict(model.eval, test)
257   sqrt(sum(abs(test$Price-yprime)^2)/length(yprime)) ##  

258   RMSE
259 }
260
261 # Evaluación de todos los modelos
262 nombre <- './Datasets Regresion/house/house'
263 cat('Error RMSE del modelo P27p4 sobre 5-fold:', mean(
264   sapply(1:5,run_k_fold,
265                     nombre,
266                     lineal.simple.P27p4), fill=T))
267 cat('Error RMSE del modelo P11p4 sobre 5-fold:', mean(
268   sapply(1:5,run_k_fold,
269                     nombre,
270                     lineal.simple.P11p4), fill=T))
271 cat('Error RMSE del modelo H13p1 sobre 5-fold:', mean(
272   sapply(1:5,run_k_fold,

```

```

266                                         nombre ,
267                                         lineal.simple.H13p1), fill=T))
268 cat('Error RMSE del modelo H40p4 sobre 5-fold:', mean(
269   sapply(1:5,run_k_fold,
270                                         nombre ,
271                                         lineal.simple.H40p4), fill=T))
272 cat('Error RMSE del modelo P16p2 sobre 5-fold:', mean(
273   sapply(1:5,run_k_fold,
274                                         nombre ,
275   lineal.simple.P16p2), fill=T))

276 # Modelo lineal compuesto por todas las variables del
277 # dataset
278 lineal.model <- lm(Price~, data=house)
279
280 summary(lineal.model)
281 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
282   model), fill=T)
283 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
284   : ', mean(sapply(1:5,run_k_fold,
285                                         nombre
286   , lineal.model), fill=T))

287 # Modelo lineal sin H2p2
288 lineal.model.fit1 <- lm(Price~.-H2p2, data=house)
289
290 summary(lineal.model.fit1)
291 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
292   model.fit1), fill=T)
293 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
294   : ', mean(sapply(1:5,run_k_fold,
295                                         nombre , lineal.model.fit1), fill=T))

296 # Modelo no lineal con P27p4 al cuadrado
297 lineal.model.fit2 <- lm(Price~.-H2p2+I(P27p4^2) , data=
298   house)
299
300 summary(lineal.model.fit2)
301 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
302   model.fit2), fill=T)
303 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
304   : ', mean(sapply(1:5,run_k_fold,
305                                         nombre , lineal.model.fit2), fill=T))

306 # Modelo no lineal con P27p4 elevado a 3
307 lineal.model.fit3 <- lm(Price~.-H2p2+I(P27p4^3) , data=
308   house)

```

```

298
299 summary(lineal.model.fit3)
300 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
     model.fit3), fill=T)
301 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
     : ', mean(sapply(1:5,run_k_fold,
302
            nombre, lineal.model.fit3), fill=T))
303
304 # Modelo no lineal con P27p4 elevado a 4
305 lineal.model.fit4 <- lm(Price~.-H2p2+I(P27p4^4), data=
     house)
306
307 summary(lineal.model.fit4)
308 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
     model.fit4), fill=T)
309 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
     : ', mean(sapply(1:5,run_k_fold,
310
            nombre, lineal.model.fit4), fill=T))
311
312 # Modelo no lineal con P11p4 elevado al cuadrado
313 lineal.model.fit5 <- lm(Price~.-H2p2+I(P27p4^2)+I(P11p4^
     2), data=house)
314
315 summary(lineal.model.fit5)
316 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
     model.fit5), fill=T)
317 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
     : ', mean(sapply(1:5,run_k_fold,
318
            nombre, lineal.model.fit5), fill=T))
319
320 # Modelo no lineal con H13p1 elevado al cuadrado
321 lineal.model.fit6 <- lm(Price~.-H2p2+I(P27p4^2)+I(P11p4^
     2)+I(H13p1^2), data=house)
322
323 summary(lineal.model.fit6)
324 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
     model.fit6), fill=T)
325 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
     : ', mean(sapply(1:5,run_k_fold,
326
            nombre, lineal.model.fit6), fill=T))
327
328 # Modelo no lineal con H13p1 elevado al cuadrado y sin
     P11p4 al cuadrado
329 lineal.model.fit7 <- lm(Price~.-H2p2+I(P27p4^2)+I(H13p1^
     2), data=house)

```

```

330
331 summary(lineal.model.fit7)
332 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
333   model.fit7), fill=T)
333 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
334   : ', mean(sapply(1:5,run_k_fold,
334
335     nombre, lineal.model.fit7), fill=T))
335
336 # Modelo no lineal con H40p4 elevado al cuadrado al
337   cuadrado
337 lineal.model.fit8 <- lm(Price~.-H2p2+I(P27p4^2)+I(H13p1^
338   2)+I(H40p4^2), data=house)
338
339 summary(lineal.model.fit8)
340 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
341   model.fit8), fill=T)
341 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
342   : ', mean(sapply(1:5,run_k_fold,
342
343     nombre, lineal.model.fit8), fill=T))
343 # Modelo no lineal con P16p2 elevado al cuadrado al
344   cuadrado
344 lineal.model.fit9 <- lm(Price~.-H2p2+I(P27p4^2)+I(H13p1^
345   2)+I(H40p4^2)+I(P16p2^2), data=house)
345
346 summary(lineal.model.fit9)
347 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
348   model.fit9), fill=T)
348 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
349   : ', mean(sapply(1:5,run_k_fold,
349
350     nombre, lineal.model.fit9), fill=T))
350
351 # Empezar con los otros
352
353 # Modelo no lineal con P1 elevado al cuadrado al
354   cuadrado
354 lineal.model.fit10 <- lm(Price~.-H2p2+I(P27p4^2)+I(H13p1^
355   2)+I(H40p4^2)+I(P16p2^2)+I(P1^2), data=house)
355
356 summary(lineal.model.fit10)
357 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
358   model.fit10), fill=T)
358 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
359   : ', mean(sapply(1:5,run_k_fold,
359
360     nombre, lineal.model.fit10), fill=T))
360

```

```

361 # Modelo no lineal con P5p1 elevado al cuadrado al
362 # cuadrado (No vale)
363 lineal.model.fit11 <- lm(Price ~ . - H2p2 + I(P27p4^2) + I(H13p1
364 ^2) + I(H40p4^2) + I(P16p2^2) +
365 I(P1^2) + I(P5p1^2), data=house
366 )
367
368 summary(lineal.model.fit11)
369 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
370 model.fit11), fill=T)
371 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
372 : ', mean(sapply(1:5, run_k_fold,
373
374 nombre, lineal.model.fit11), fill=T))
375
376 # Modelo no lineal con P11p4 elevado al cuadrado al
377 # cuadrado
378 lineal.model.fit12 <- lm(Price ~ . - H2p2 + I(P27p4^2) + I(H13p1
379 ^2) + I(H40p4^2) + I(P16p2^2) +
380 I(P1^2) + I(P11p4^2), data=
381 house)
382
383 summary(lineal.model.fit12)
384 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
385 model.fit12), fill=T)
386 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
387 : ', mean(sapply(1:5, run_k_fold,
388
389 nombre, lineal.model.fit12), fill=T))
390
391 # Modelo no lineal con P6p2 elevado al cuadrado al
392 # cuadrado (No vale)
393 lineal.model.fit13 <- lm(Price ~ . - H2p2 + I(P27p4^2) + I(H13p1
394 ^2) + I(H40p4^2) + I(P16p2^2) +
395 I(P1^2) + I(P11p4^2) + I(P6p2^2),
396 data=house)
397
398 summary(lineal.model.fit13)
399 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
400 model.fit13), fill=T)
401 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
402 : ', mean(sapply(1:5, run_k_fold,
403
404 nombre, lineal.model.fit13), fill=T))
405
406 # Modelo no lineal con H8p2 elevado al cuadrado
407 lineal.model.fit14 <- lm(Price ~ . - H2p2 + I(P27p4^2) + I(H13p1
408 ^2) + I(H40p4^2) + I(P16p2^2) +

```

```

390                               I(P1^2)+I(P11p4^2)+I(H8p2^2) ,
391   data=house)
392 summary(lineal.model.fit14)
393 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
394 model.fit14), fill=T)
394 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
395 : ', mean(sapply(1:5,run_k_fold,
396
397           nombre, lineal.model.fit14), fill=T))
398
399 # Modelo no lineal con P18p2
400 lineal.model.fit15 <- lm(Price~.-H2p2+I(P27p4^2)+I(H13p1
401 ^2)+I(H40p4^2)+I(P16p2^2)+
402                               I(P1^2)+I(P11p4^2)+I(P18p2^2)
403 , data=house) #I(H8p2^2)+
404
405 summary(lineal.model.fit15)
406 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
407 model.fit15), fill=T)
408 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
409 : ', mean(sapply(1:5,run_k_fold,
410
411           nombre, lineal.model.fit15), fill=T))
412
413 # Modelo no lineal añadiendo H10p1
414 lineal.model.fit16 <- lm(Price~.-H2p2+I(P27p4^2)+I(H13p1
415 ^2)+I(H40p4^2)+I(P16p2^2)+
416                               I(P1^2)+I(P11p4^2)+I(P18p2^2)
417 +I(H10p1^2) , data=house)
418
419 summary(lineal.model.fit16)
420 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
421 model.fit16), fill=T)
422 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
423 : ', mean(sapply(1:5,run_k_fold,
424
425           nombre, lineal.model.fit16), fill=T))
426
427 # Modelo no lineal añadiendo H18pA al cuadrado
428 lineal.model.fit17 <- lm(Price~.-H2p2+I(P27p4^2)+I(H13p1
429 ^2)+I(H40p4^2)+I(P16p2^2)+
430                               I(P1^2)+I(P11p4^2)+I(P18p2^2)
431 +I(H10p1^2) +
432                               I(H18pA^2) , data=house)
433
434 summary(lineal.model.fit17)
435 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
436 model.fit17), fill=T)

```

```

422 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
423   : ', mean(sapply(1:5,run_k_fold,
424
425     nombre, lineal.model.fit17), fill=T))
426 # Modelo no lineal añadiendo H13p1 al cuadrado
427 lineal.model.fit18 <- lm(Price~.-H2p2+I(P27p4^2)+I(H13p1
428   ^2)+I(H40p4^2)+I(P16p2^2) +
429   I(P1^2)+I(P11p4^2)+I(P18p2^2)
430   +I(H10p1^2) +
431   I(H18pA^2)+I(H13p1^2), data=
432   house)
433
434 summary(lineal.model.fit18)
435 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
436   model.fit18), fill=T)
437 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
438   : ', mean(sapply(1:5,run_k_fold,
439
440     nombre, lineal.model.fit18), fill=T))
441
442 # Modelo no lineal añadiendo P14p9 al cuadrado
443 lineal.model.fit19 <- lm(Price~.-H2p2+I(P27p4^2)+I(H13p1
444   ^2)+I(H40p4^2)+I(P16p2^2) +
445   I(P1^2)+I(P11p4^2)+I(P18p2^2)
446   +I(H10p1^2) +
447   I(H18pA^2)+I(P14p9^2), data=
448   house)
449
450 summary(lineal.model.fit19)
451 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
452   model.fit19), fill=T)
453 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
454   : ', mean(sapply(1:5,run_k_fold,
455
456     nombre, lineal.model.fit19), fill=T))
457
458 # Modelo no lineal con raíz de P18p2
459 lineal.model.fit20 <- lm(Price~.-H2p2+I(P27p4^2)+I(H13p1
460   ^2)+I(H40p4^2)+I(P16p2^2) +
461   I(P1^2)+I(P11p4^2)+I(P18p2^2)
462   +I(sqrt(P18p2))+I(H10p1^2) +
463   I(H18pA^2), data=house)
464
465 summary(lineal.model.fit20)
466 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
467   model.fit20), fill=T)
468 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
469   : ', mean(sapply(1:5,run_k_fold,
470
471     nombre, lineal.model.fit20), fill=T))

```

```

452
453 # Modelo no lineal con raíz de H18pA
454 lineal.model.fit21 <- lm(Price ~ . - H2p2 + I(P27p4^2) + I(H13p1
455 ^2) + I(H40p4^2) + I(P16p2^2) +
456 I(P1^2) + I(P11p4^2) + I(P18p2^2)
457 + I(H10p1^2) +
458 I(H18pA^2) + I(sqrt(P18p2)) +
459 sqrt(H18pA), data=house)
460 summary(lineal.model.fit21)
461 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
462 model.fit21), fill=T)
463 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
464 : ', mean(sapply(1:5, run_k_fold,
465
466 nombre, lineal.model.fit21), fill=T))
467
468 # Modelo no lineal con raíz de P15p3
469 lineal.model.fit22 <- lm(Price ~ . - H2p2 + I(P27p4^2) + I(H13p1
470 ^2) + I(H40p4^2) + I(P16p2^2) +
471 I(P1^2) + I(P11p4^2) + I(P18p2^2)
472 + I(H10p1^2) +
473 I(H18pA^2) + I(sqrt(P18p2)) +
474 sqrt(H18pA) + sqrt(P15p3), data=house)
475 summary(lineal.model.fit22)
476 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
477 model.fit22), fill=T)
478 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
479 : ', mean(sapply(1:5, run_k_fold,
480
481 nombre, lineal.model.fit22), fill=T))
482
483 # Modelo no lineal con raíz de P1
484 lineal.model.fit23 <- lm(Price ~ . - H2p2 + I(P27p4^2) + I(H13p1
485 ^2) + I(H40p4^2) + I(P16p2^2) +
486 I(P1^2) + I(P11p4^2) + I(P18p2^2)
487 + I(H10p1^2) +
488 I(H18pA^2) + I(sqrt(P18p2)) +
489 sqrt(H18pA) + sqrt(P15p3) + sqrt(P1), data=house)
490 summary(lineal.model.fit23)
491 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
492 model.fit23), fill=T)
493 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
494 : ', mean(sapply(1:5, run_k_fold,
495
496 nombre, lineal.model.fit23), fill=T))
497
498 # Modelo no lineal con H10p1 elevado a 4
499 lineal.model.fit24 <- lm(Price ~ . - H2p2 + I(P27p4^2) + I(H13p1
500 ^2) + I(H40p4^2) + I(P16p2^2) +

```

```

481                               I(P1^2)+I(P11p4^2)+I(P18p2^2)
482                               +I(H10p1^4) +
483                               I(H18pA^2)+I(sqrt(P18p2))+
484 summary(lineal.model.fit24)
485 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
486 model.fit24), fill=T)
487 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
488 : ', mean(sapply(1:5,run_k_fold,
489
490                               nombre, lineal.model.fit24), fill=T))
491
492 lineal.model.fit25 <- lm(Price~.-H2p2+I(P27p4^2)+I(H13p1
493 ^2)+I(H40p4^2)+I(P16p2^2) +
494                               I(P1^2)+I(P11p4^2)+I(P18p2^2)
495                               +I(H10p1^4) +
496                               I(H18pA^2)+I(sqrt(P18p2))+
497                               sqrt(H18pA)+sqrt(P15p3)+
498                               I(sqrt(P1*P15p3)), data=house
499 )
500 summary(lineal.model.fit25)
501 cat('RMSE del modelo: ', evaluate.rmse.house(lineal.
502 model.fit25), fill=T)
503 cat('Error RMSE del modelo lineal compuesto sobre 5-fold
504 : ', mean(sapply(1:5,run_k_fold,
505
506                               nombre, lineal.model.fit25), fill=T))
507
508 #### Modelos k-NN
509
510 # Definición de la función para evaluar un modelo KNN
511 evaluate.knn_rmse.house <- function(model) {
512   yprime = model$fitted.values
513   sqrt(sum(abs(house$Price-yprime)^2)/length(yprime)) #
514   Calcular RMSE
515 }
516
517 # Evaluar los modelos mediante validación cruzada 5-fold
518 run_knn_k_fold_house <- function(i, x, model,
519   standarize=NULL, tt = "test") {
520   # Cargar conjuntos de entrenamiento
521   file <- paste(x, "-5-", i, "tra.dat", sep=" ");
522   x_tra <- read.csv(file, comment.char="@")
523
524   # Cargar conjuntos de test
525   file <- paste(x, "-5-", i, "tst.dat", sep=" ");
526   x_tst <- read.csv(file, comment.char="@")
527
528   names(x_tra) <- c('P1', 'P5p1', 'P6p2', 'P11p4', '
529   P14p9', 'P15p1', 'P15p3',

```

```

515             'P16p2', 'P18p2', 'P27p4', 'H2p2', '
516             'H8p2', 'H10p1', 'H13p1',
517             'H18pA', 'H40p4', 'Price')
518 names(x_tst) <- names(x_tra)
519
520 # Estandarizar las variables proporcionadas
521 for (i in standarize) {
522   # Aplicar sobre x_tra
523   min.v <- min(x_tra[i])
524   max.v <- max(x_tra[i])
525
526   x_tra[[paste(i, '_rescaled', sep= '')]] <- unlist((x_
527   tra[i] - min.v)/(max.v-min.v))
528
529   # Aplicar sobre x_tst
530   min.v <- min(x_tst[i])
531   max.v <- max(x_tst[i])
532
533   x_tst[[paste(i, '_rescaled', sep= '')]] <- unlist((x_
534   tst[i] - min.v)/(max.v-min.v))
535 }
536
537 if (tt == "train") { test <- x_tra }
538 else { test <- x_tst }
539
540 # Entrenar el modelo sobre el conjunto de train
541 form <- terms(model)
542 model.eval <- kknn(formula=form, train=x_tra, test=
543   test, k=ncol(model$CL),
544   distance=model$distance)
545
546 # Evaluación del RMSE sobre test
547 yprime=model.eval$fitted.values
548 sqrt(sum(abs(test$Price-yprime)^2)/length(yprime)) ###
549 RMSE
550
551 # Modelo por defecto
552 knn.model.fit1 <- kknn(Price~, house, house)
553
554 summary(knn.model.fit1)
555 cat('RMSE del modelo: ', evaluate.knn_rmse.house(knn.
556   model.fit1), fill=T)
557 cat('Error RMSE del modelo kknn compuesto sobre 5-fold:'
558   , mean(sapply(1:5, run_knn_k_fold_house,
559
560   nombre, knn.model.fit1), fill=T))
561
562 # Reescalar P1 al intervalo [0,1]

```

```

556 min.p1 <- min(house$P1)
557 max.p1 <- max(house$P1)
558
559 house$P1_rescaled = (house$P1-min.p1)/(max.p1-min.p1)
560
561 # Entrenar otro modelo con P1 reescalado
562 knn.model.fit2 <- kknn(Price~.-P1, house, house)
563
564 summary(knn.model.fit2)
565 cat('RMSE del modelo: ', evaluate.knn_rmse.house(knn.
      model.fit2), fill=T)
566 cat('Error RMSE del modelo kknn compuesto sobre 5-fold: '
      , mean(sapply(1:5,run_knn_k_fold_house,
567
      nombre, knn.model.fit2, 'P1')), fill=T))
568
569 # Entrenar otro modelo con transformaciones para
      normalizar las variables
570 knn.model.fit3 <- kknn(Price~sqrt(P1_rescaled)+P5p1+I(
      P6p2^(1/3))+sqrt(P11p4)+sqrt(P14p9)+
      I(P15p1^2)+I(P15p3^(1/3))+I(
      P16p2^2)+sqrt(P18p2)+sqrt(P27p4)+
      sqrt(H2p2)+sqrt(H8p2)+I(H10p1^
      4)+sqrt(H13p1)+
      sqrt(H18pA)+H40p4, house, house
      )
571
572 summary(knn.model.fit3)
573 cat('RMSE del modelo: ', evaluate.knn_rmse.house(knn.
      model.fit3), fill=T)
574 cat('Error RMSE del modelo kknn compuesto sobre 5-fold: '
      , mean(sapply(1:5,run_knn_k_fold_house,
575
      nombre, knn.model.fit3, 'P1')), fill=T))
576
577 # Entrenar otro modelo con transformaciones para
      normalizar las variables, pero sin reescalar P1
578 knn.model.fit4 <- kknn(Price~sqrt(P1)+P5p1+I(P6p2^(1/3))+
      +sqrt(P11p4)+sqrt(P14p9)+
      I(P15p1^2)+I(P15p3^(1/3))+I(
      P16p2^2)+sqrt(P18p2)+sqrt(P27p4)+
      sqrt(H2p2)+sqrt(H8p2)+I(H10p1^
      4)+sqrt(H13p1)+
      sqrt(H18pA)+H40p4, house, house
      )
579
580 summary(knn.model.fit4)
581 cat('RMSE del modelo: ', evaluate.knn_rmse.house(knn.
      model.fit3), fill=T)

```

```

588 cat('Error RMSE del modelo kknn compuesto sobre 5-fold:'
      , mean(sapply(1:5, run_knn_k_fold_house,
589           nombre, knn.model.fit4), fill=T))
590
591 # Evaluar el mejor valor de k
592 evaluate.knn <- function(k, model, standarizate=NULL) {
593   # Se obtiene la formula del modelo
594   form <- terms(model)
595
596   # Evaluar el modelo
597   model.eval <- kknn(form, house, house, k=k)
598   meausures <- c(k, evaluate.knn_rmse.house(model.eval),
599                   mean(sapply(1:5, run_knn_k_fold_house,
600                         nombre, model.eval, standarizate)))
601   names(meausures) <- c('k', 'RMSE', 'RMSE 5-fold')
602
603   return(meausures)
604 }
605
606 # Valores de K usados en las pruebas
607 k.values <- c(1,3,7,13,21,51,75,103)
608
609 # Modelos evaluados sobre knn.model.fit4
610 normal.model.metrics <- sapply(k.values, evaluate.knn,
611   knn.model.fit4)
611 normal.model.metrics
612
613 # Modelos evaluados sobre knn.model.fit1
614 original.model.metrics <- sapply(k.values, evaluate.knn,
615   knn.model.fit1)
615 original.model.metrics

```

C. Script3.R

Este *script* fue usado para la realización de todos los tests estadísticos usados para la comparación de métodos.

```

1 #
# ##########
2 # Nombre del script: script3.R
3 # Desarrollado por: Nicolás Cubero Torres
4 # Descripción: Script para el desarrollo de tests estadísticos para realizar
5 # la comparación de modelos
6 # Nota: Este script ha sido desarrollado para ejecutarse de forma interactiva

```

```

7 #           línea por línea
8 #
9 ##########
10 # Script para la comparación de modelos
11 results.train <- read.csv('./regr_train_alumnos.csv',
12   row.names = 1)
13 results.test <- read.csv('./regr_test_alumnos.csv', row.
14   names = 1)
15
16 # Sobre el conjunto de train
17 # Comparar out_train_lm con out_train_kknn (referencia)
18   con Wilcoxon
19 difs <- (results.train[,1] - results.train[,2]) /
20   results.train[,1]
21 wilc_1_2 <- cbind(ifelse (difs<0, abs(difs)+0.1, 0+0.1),
22   ifelse (difs>0, abs(difs)+0.1, 0+0.1))
23 colnames(wilc_1_2) <- c(colnames(results.train)[1],
24   colnames(results.train)[2])
25 head(wilc_1_2)
26
27 # Aplicar test y calcular R+ y R-
28 LMvsKNNTst <- wilcox.test(wilc_1_2[,1], wilc_1_2[,2],
29   alternative = "two.sided", paired=TRUE)
30 Rmas <- LMvsKNNTst$statistic
31 pvalue <- LMvsKNNTst$p.value
32
33 LMvsKNNTst <- wilcox.test(wilc_1_2[,2], wilc_1_2[,1],
34   alternative = "two.sided", paired=TRUE)
35 Rmenos <- LMvsKNNTst$statistic
36
37 cat('Test modelo lineal (R+) vs modelo k-NN:(R-)', fill=
38   T)
39 cat('Valor R+: ',Rmas, fill=T)
40 cat('Valor R-: ',Rmenos, fill=T)
41 cat('p-value del test: ',pvalue, fill=T)
42
43 # Sobre el conjunto de test
44 # Comparar out_train_lm con out_train_kknn (referencia)
45   con Wilcoxon
46 difs <- (results.test[,1] - results.test[,2]) / results.
47   test[,1]
48 wilc_1_2 <- cbind(ifelse (difs<0, abs(difs)+0.1, 0+0.1),
49   ifelse (difs>0, abs(difs)+0.1, 0+0.1))
50 colnames(wilc_1_2) <- c(colnames(results.test)[1],
51   colnames(results.test)[2])
52 head(wilc_1_2)
53
54 # Aplicar test y calcular R+ y R-

```

```

43 LMvsKNNtst <- wilcox.test(wilc_1_2[,1], wilc_1_2[,2],
44   alternative = "two.sided", paired=TRUE)
45 Rmas <- LMvsKNNtst$statistic
46 pvalue <- LMvsKNNtst$p.value
47
48 LMvsKNNtst <- wilcox.test(wilc_1_2[,2], wilc_1_2[,1],
49   alternative = "two.sided", paired=TRUE)
50 Rmenos <- LMvsKNNtst$statistic
51
52 cat('Test modelo lineal (R+) vs modelo k-NN:(R-)', fill=
53   T)
54 cat('Valor R+: ',Rmas, fill=T)
55 cat('Valor R-: ',Rmenos, fill=T)
56 cat('p-value del test: ',pvalue, fill=T)
57
58 # Aplicar el test de Friedman sobre el conjunto de
59 # entrenamiento
60 test_friedman <- friedman.test(as.matrix(results.train))
61 test_friedman
62
63 # Aplicar el test post-hoc de Holm para averiguar qué
64 # par es diferente
65 tam <- dim(results.train)
66 groups <- rep(1:tam[2], each=tam[1])
67 pairwise.wilcox.test(as.matrix(results.train), groups, p
68   .adjust = "holm", paired = TRUE)
69
70 # Aplicar el test de Friedman sobre el conjunto de test
71 test_friedman <- friedman.test(as.matrix(results.test))
72 test_friedman
73
74 # Aplicar el test post-hoc de Holm para averiguar qué
75 # par es diferente
76 tam <- dim(results.test)
77 groups <- rep(1:tam[2], each=tam[1])
78 pairwise.wilcox.test(as.matrix(results.train), groups, p
    .adjust = "holm", paired = TRUE)

```

D. Script4.R

Este *script* fue usado para la realización del análisis exploratorio de datos sobre el *dataset vehicle*.

```

1 #
# ##########
2 # Nombre del script: script4.R

```

```

3 # Desarrollado por: Nicolás Cubero Torres
4 # Descripción: Script para el análisis exploratorio del
5 # dataset vehicle
6 #
7 # Nota: Este script ha sido desarrollado para ejecutarse
8 # de forma interactiva
9 # línea por línea
10 #
11 #####
12 #
13 # Librerías cargadas
14 library('moments')
15 library('ggplot2')
16 library('dplyr')
17 library('GGally')
18 library('corrplot')
19 #
20 #
21 # Función para leer el dataset vehicle
22 read.vehicle.dataset <- function(filename) {
23   # Cargar datos
24   dat <- read.table(filename, comment.char="@", sep=',')
25   #
26   # Asignar nombres
27   names(dat) <- c('Compactness', 'Circularity', ,
28     'Distance_circularity',
29     'Radius_ratio', 'Praxis_aspect_ratio',
30     'Max_length_aspect_ratio', 'Scatter_
31     ratio',
32     'Elongatedness', 'Praxis_rectangular',
33     'Length_rectangular',
34     'Major_variance', 'Minor_variance', ,
35     'Gyration_radius',
36     'Major_skewness', 'Minor_skewness', ,
37     'Minor_kurtosis',
38     'Major_kurtosis', 'Hollows_ratio', ,
39     'Class')
40   #
41   return(dat)
42 }
43 #
44 # Cargar el dataset vehicle
45 vehicle <- read.vehicle.dataset('./Datasets
46   Clasificacion/vehicle/vehicle.dat')
47 #
48 # Comprobar la existencia de Missing Values
49 any(is.na(vehicle))
50 #
51 # Analizar brevemente su estructura

```

```

41 str(vehicle)
42 head(vehicle)
43
44 # Determinar las distribuciones de las variables
45 # Determinar los estadísticos de posición:
46 # Valores mínimo y máximos, media, 1er cuartil, mediana,
47 # 3er cuartil
48 cat('Estadísticos de posición: Valores mínimo y máximos,
49     media, 1er cuartil,',
50     ' mediana, 3er cuartil', fill=T)
51 summary(vehicle[,c(1:ncol(vehicle)-1)])
52
53 # Determinar los estadísticos de dispersión: desviación
54 # típica
55 cat('Desviación típica de los atributos', fill=T)
56 apply(vehicle[,c(1:ncol(vehicle)-1)], MARGIN=2, FUN=sd)
57
58 # Determinar coeficientes de Skew y Kurtosis
59 cat('Coeficientes de Skew y Kurtosis', fill=T)
60 skew_kurtosis <- apply(vehicle[,c(1:ncol(vehicle)-1)],
61                         MARGIN=2,
62                         FUN=function(x) {c(skewness(x),
63                                     kurtosis(x))})
63 rownames(skew_kurtosis) <- c('Skew', 'kurtosis')
64 skew_kurtosis
65
66 # Analizar las distribuciones de las variables
67 # Compactness
68 # Diagrama boxplot
69 graf.compactness.boxplot <- ggplot(vehicle, aes(x='
70     Compactness', y=Compactness)) +
71     geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
72     outlier.shape=8) +
73     labs(cation='center', y=' ', x=' ')
74 graf.compactness.boxplot
75
76 # Histograma
77 graf.compactness.histogram <- ggplot(vehicle, aes(
78     Compactness, fill=Class)) +
79     geom_histogram(binwidth=1, color='black', alpha=0.4) +
80     labs(y='frecuencia absoluta')
81 graf.compactness.histogram
82
83 # Estudiamos normalidad con el test de Shapiro-Milk
84 shapiro.test(vehicle$Compactness)
85
86 # Circularity
87 # Diagrama boxplot

```

```

81 graf.circularity.boxplot <- ggplot(vehicle, aes(x='
82   Circularity', y=Circularity)) +
83   geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
84     outlier.shape=8) +
85   labs(cation='center', y='', x='')
86 graf.circularity.boxplot
87
88 # Histograma
89 graf.circularity.histogram <- ggplot(vehicle, aes(
90   Circularity, fill=Class)) +
91   geom_histogram(binwidth=1, color='black', alpha=0.4) +
92   labs(y='frecuencia absoluta')
93 graf.circularity.histogram
94
95 # Estudiamos normalidad con el test de Shapiro-Milk
96 shapiro.test(vehicle$Circularity)
97
98 # Distance_circularity
99 # Diagrama boxplot
100 graf.distance_circularity.boxplot <- ggplot(vehicle, aes(
101   x='Distance_circularity',
102                           y=
103   Distance_circularity)) +
104   geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
105     outlier.shape=8) +
106   labs(cation='center', y='', x='')
107 graf.distance_circularity.boxplot
108
109 # Histograma
110 graf.distance_circularity.histogram <- ggplot(vehicle,
111                                               aes(Distance_
112         circularity, fill=Class)) +
113   geom_histogram(binwidth=1, color='black', alpha=0.4) +
114   labs(y='frecuencia absoluta')
115 graf.distance_circularity.histogram
116
117 # Estudiamos normalidad con el test de Shapiro-Milk
118 shapiro.test(vehicle$Distance_circularity)
119
120 # Radius_ratio
121 # Diagrama boxplot
122 graf.radius_ratio.boxplot <- ggplot(vehicle, aes(x='
123   Radius_ratio',
124                           y=Radius_ratio)) +
125   geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
126     outlier.shape=8) +
127   labs(cation='center', y='', x='')
128 graf.radius_ratio.boxplot
129

```

```

120
121 # Histograma
122 graf.radius_ratio.histogram <- ggplot(vehicle,
123                                         aes(Radius
124                                         _ratio, fill=Class)) +
125                                         geom_histogram(binwidth=3, color='black', alpha=0.4) +
126                                         labs(y='frecuencia absoluta')
127 graf.radius_ratio.histogram
128
129 # Calcular rango de valores pertenecientes a la
130 # distribución
131 iqr.radius_ratio <- IQR(vehicle$Radius_ratio)
132 quantiles.radius_ratio <- quantile(vehicle$Radius_ratio
133                                         )
134
135 q1.radius_ratio <- quantiles.radius_ratio[2]
136 q3.radius_ratio <- quantiles.radius_ratio[4]
137
138 cat('Intervalo de distribución: [', q1.radius_ratio-1.5
139 *iqr.radius_ratio, ',',
140     q3.radius_ratio+1.5*iqr.radius_ratio, ']', fill=T)
141
142 # Aislamos los outliers que se encuentran por encima del
143 # límite superior
144 vehicle.outliers.radius_ratio <- vehicle %>% filter(
145     Radius_ratio>276)
146 vehicle.outliers.radius_ratio %>% select(Radius_ratio,
147                                               Class)
148
149 # Estudiamos normalidad con el test de Shapiro-Milk
150 shapiro.test(vehicle$Radius_ratio)
151
152 # Praxis_aspect_ratio
153 # Diagrama boxplot
154 pdf('Praxis_aspect_ratio_boxplot.pdf')
155 graf.praxis_aspect_ratio.boxplot <- ggplot(vehicle,
156                                         aes(x=
157                                         Praxis_aspect_ratio,
158                                         y=
159                                         Praxis_aspect_ratio)) +
160                                         geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
161                                         outlier.shape=8) +
162                                         labs(caption='center', y=' ', x=' ')
163 graf.praxis_aspect_ratio.boxplot
164 dev.off()
165
166 # Histograma
167 pdf('Praxis_aspect_ratio_histograma.pdf')
168 graf.praxis_aspect_ratio.histogram <- ggplot(vehicle,

```

```

159                                     aes(
160             Praxis_aspect_ratio, fill=Class)) +
161             geom_histogram(binwidth=1, color='black', alpha=0.4) +
162             labs(y='frecuencia absoluta')
163 graf.praxis_aspect_ratio.histogram
164 dev.off()
165
166 # Max_length_aspect_ratio
167 # Diagrama boxplot
168 graf.max_length_aspect_ratio.boxplot <- ggplot(vehicle,
169                                         aes(x='Max_
170                                         _length_aspect_ratio',
171                                         y=Max_
172                                         _length_aspect_ratio)) +
173             geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
174                         outlier.shape=8) +
175             labs(cation='center', y='', x='')
176 graf.max_length_aspect_ratio.boxplot
177
178 # Histograma
179 graf.max_length_aspect_ratio.histogram <- ggplot(vehicle
180 ,
181                                         aes(Max_length_
182                                         aspect_ratio, fill=Class)) +
183             geom_histogram(binwidth=1, color='black', alpha=0.4) +
184             labs(y='frecuencia absoluta')
185 graf.max_length_aspect_ratio.histogram
186
187 # Scatter_ratio
188 # Diagrama boxplot
189 pdf('Scatter_ratio_boxplot.pdf')
190 graf.scatter_ratio.boxplot <- ggplot(vehicle, aes(x='
191                                         Scatter_ratio',
192                                         y=
193                                         Scatter_ratio)) +
194             geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
195                         outlier.shape=8) +
196             labs(cation='center', y='', x='')
197 graf.scatter_ratio.boxplot
198 dev.off()
199
200 # Histograma
201 pdf('Scatter_ratio_histograma.pdf')
202 graf.scatter_ratio.histogram <- ggplot(vehicle, aes(
203                                         Scatter_ratio,
204                                         fill
205                                         =Class)) +
206             geom_histogram(binwidth=2, color='black', alpha=0.4) +
207             labs(y='frecuencia absoluta')

```

```

197 graf.scatter_ratio.histogram
198 dev.off()
199
200 # Elongatedness
201 # Diagrama boxplot
202 pdf('Elongatedness_boxplot.pdf')
203 graf.elongatedness.boxplot <- ggplot(vehicle, aes(x=
204     Elongatedness,
205             y=
206     Elongatedness)) +
207     geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
208     outlier.shape=8) +
209     labs(cation='center', y='', x='')
210 graf.elongatedness.boxplot
211 dev.off()
212
213 # Histograma
214 pdf('Elongatedness_histograma.pdf')
215 graf.elongatedness.histogram <- ggplot(vehicle, aes(
216     Elongatedness,
217             fill
218     =Class)) +
219     geom_histogram(binwidth=1, color='black', alpha=0.4) +
220     labs(y='frecuencia absoluta')
221 graf.elongatedness.histogram
222 dev.off()
223
224 # Praxis_rectangular
225 # Diagrama boxplot
226 pdf('Praxis_rectangular_boxplot.pdf')
227 graf.praxis_rectangular.boxplot <- ggplot(vehicle, aes(x
228     ='Praxis_rectangular',
229             y=
230     Praxis_rectangular)) +
231     geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
232     outlier.shape=8) +
233     labs(cation='center', y='', x='')
234 graf.praxis_rectangular.boxplot
235 dev.off()
236
237 # Histograma
238 pdf('Praxis_rectangular_histograma.pdf')
239 graf.praxis_rectangular.histogram <- ggplot(vehicle, aes(
240     Praxis_rectangular,
241             fill
242     =Class)) +
243     geom_histogram(binwidth=1, color='black', alpha=0.4) +
244     labs(y='frecuencia absoluta')
245 graf.praxis_rectangular.histogram

```

```

236 dev.off()
237
238 # Length_rectangular
239 # Diagrama boxplot
240 pdf('Length_rectangular_boxplot.pdf')
241 graf.length_rectangular.boxplot <- ggplot(vehicle, aes(x
242 = 'Length_rectangular',
243
244 = Length_rectangular)) +
245 geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
246 outlier.shape=8) +
247 labs(cation='center', y='', x='')
248 graf.length_rectangular.boxplot
249 dev.off()
250
251 # Histograma
252 pdf('Length_rectangular_histograma.pdf')
253 graf.length_rectangular.histogram <- ggplot(vehicle, aes(
254 Length_rectangular,
255
256 fill=Class)) +
257 geom_histogram(binwidth=1, color='black', alpha=0.4) +
258 labs(y='frecuencia absoluta')
259 graf.length_rectangular.histogram
260 dev.off()
261
262 # Major_variance
263 # Diagrama boxplot
264 pdf('Major_variance_boxplot.pdf')
265 graf.major_variance.boxplot <- ggplot(vehicle, aes(x=
266 'Major_variance',
267
268 = Major_variance)) +
269 geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
270 outlier.shape=8) +
271 labs(cation='center', y='', x='')
272 graf.major_variance.boxplot
273 dev.off()
274
275 # Histograma
276 pdf('Major_variance_histograma.pdf')
277 graf.major_variance.histogram <- ggplot(vehicle, aes(
278 Major_variance,
279
280 fill=Class)) +
281 geom_histogram(binwidth=3, color='black', alpha=0.4) +
282 labs(y='frecuencia absoluta')
283 graf.major_variance.histogram
284 dev.off()

```

```

275 # Analizar outliers por encima del valor 300
276 vehicle.outliers.major_variance <- vehicle %>% filter(
277   Major_variance>300)
278 vehicle.outliers.major_variance %>% select(Major_
279   variance, Class)
280
281 # Minor_variance
282 # Diagrama boxplot
283 pdf('Minor_variance_boxplot.pdf')
284 graf.minor_variance.boxplot <- ggplot(vehicle, aes(x='
285   Minor_variance',
286   y=
287   Minor_variance)) +
288   geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
289     outlier.shape=8) +
290   labs(caption='center', y=' ', x=' ')
291 graf.minor_variance.boxplot
292 dev.off()
293
294 # Histograma
295 pdf('Minor_variance_histograma.pdf')
296 graf.minor_variance.histogram <- ggplot(vehicle, aes(
297   Minor_variance,
298   fill=Class)) +
299   geom_histogram(binwidth=15, color='black', alpha=0.4)
300   +
301   labs(y='frecuencia absoluta')
302 graf.minor_variance.histogram
303 dev.off()
304
305 # Gyration_radius
306 # Diagrama boxplot
307 pdf('Gyration_radius_boxplot.pdf')
308 graf.gyration_radius.boxplot <- ggplot(vehicle, aes(x='
309   Gyration_radius',
310   y=
311   Gyration_radius)) +
312   geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
313     outlier.shape=8) +
314   labs(caption='center', y=' ', x=' ')
315 graf.gyration_radius.boxplot
316 dev.off()
317
318 # Histograma
319 pdf('Gyration_radius_histograma.pdf')
320 graf.gyration_radius.histogram <- ggplot(vehicle, aes(
321   Gyration_radius,

```

```

312     fill=Class)) +
313     geom_histogram(binwidth=3, color='black', alpha=0.4) +
314     labs(y='frecuencia absoluta')
315 graf.gyration_radius.histogram
316 dev.off()
317
318 # Major_skewness
319 # Diagrama boxplot
320 pdf('Major_skewness_boxplot.pdf')
321 graf.major_skewness.boxplot <- ggplot(vehicle, aes(x=
322     Major_skewness',
323             y=
324     Major_skewness)) +
325     geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
326                 outlier.shape=8) +
327     labs(caption='center', y='', x='')
328 graf.major_skewness.boxplot
329 dev.off()
330
331 # Histograma
332 pdf('Major_skewness_histограмa.pdf')
333 graf.major_skewness.histogram <- ggplot(vehicle, aes(
334     Major_skewness,
335             fill=Class)) +
336     geom_histogram(binwidth=2, color='black', alpha=0.4) +
337     labs(y='frecuencia absoluta')
338 graf.major_skewness.histogram
339 dev.off()
340
341 # Minor_skewness
342 # Diagrama boxplot
343 pdf('Minor_skewness_boxplot.pdf')
344 graf.minor_skewness.boxplot <- ggplot(vehicle, aes(x=
345     Minor_skewness',
346             y=
347     Minor_skewness)) +
348     geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
349                 outlier.shape=8) +
350     labs(caption='center', y='', x='')
351 graf.minor_skewness.boxplot
352 dev.off()
353
354 # Histograma
355 pdf('Minor_skewness_histограма.pdf')
356 graf.minor_skewness.histogram <- ggplot(vehicle, aes(
357     Minor_skewness,

```

```

350     fill=Class)) +
351     geom_histogram(binwidth=1, color='black', alpha=0.4) +
352     labs(y='frecuencia absoluta')
353 graf.minor_skewness.histogram
354 dev.off()

355
356 # Minor_kurtosis
357 # Diagrama boxplot
358 pdf('Minor_kurtosis_boxplot.pdf')
359 graf.minor_kurtosis.boxplot <- ggplot(vehicle, aes(x=
360     'Minor_kurtosis',
361             y=
362     'Minor_kurtosis)) +
363     geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
364     outlier.shape=8) +
365     labs(cation='center', y=' ', x=' ')
366 graf.minor_kurtosis.boxplot
367 dev.off()

368 # Histograma
369 pdf('Minor_kurtosis_histограма.pdf')
370 graf.minor_kurtosis.histogram <- ggplot(vehicle, aes(
371     'Minor_kurtosis',
372             fill=Class)) +
373     geom_histogram(binwidth=1, color='black', alpha=0.4) +
374     labs(y='frecuencia absoluta')
375 graf.minor_kurtosis.histogram
376 dev.off()

377 # Major_kurtosis
378 # Diagrama boxplot
379 pdf('Major_kurtosis_boxplot.pdf')
380 graf.major_kurtosis.boxplot <- ggplot(vehicle, aes(x=
381     'Major_kurtosis',
382             y=
383     'Major_kurtosis)) +
384     geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
385     outlier.shape=8) +
386     labs(cation='center', y=' ', x=' ')
387 graf.major_kurtosis.boxplot
388 dev.off()

389 # Histograma
390 pdf('Major_kurtosis_histограма.pdf')
391 graf.major_kurtosis.histogram <- ggplot(vehicle, aes(
392     'Major_kurtosis',

```

```

388     fill=Class)) +
389     geom_histogram(binwidth=1, color='black', alpha=0.4) +
390     labs(y='frecuencia absoluta')
391 graf.major_kurtosis.histogram
392 dev.off()
393
394 # Major_kurtosis
395 # Diagrama boxplot
396 pdf('Hollows_ratio_boxplot.pdf')
397 graf.hollows_ratio.boxplot <- ggplot(vehicle, aes(x=
398     Hollows_ratio,
399             y=
400     Hollows_ratio)) +
401     geom_boxplot(outlier.alpha=0.4, outlier.colour='red',
402     outlier.shape=8) +
403     labs(caption='center', y='', x='')
404 graf.hollows_ratio.boxplot
405 dev.off()
406
407 # Histograma
408 pdf('Hollows_ratio_histograma.pdf')
409 graf.hollows_ratio.histogram <- ggplot(vehicle, aes(
410     Hollows_ratio,
411             fill=Class)) +
412     geom_histogram(binwidth=1, color='black', alpha=0.4) +
413     labs(y='frecuencia absoluta')
414 graf.hollows_ratio.histogram
415 dev.off()
416
417 # Class
418 # Analizar la distribución de los valores
419 table(vehicle$Class)
420
421 # Representar gráficamente esta información
422 pdf('Class_barplot.pdf')
423 graf.class <- ggplot(vehicle, aes(x=Class, fill=Class)) +
424     geom_bar()
425 graf.class
426 dev.off()
427
428 # Representar diagramas scatterplot entre los pares de
429 # variables
430 png('relaciones_variables_vehicle.png', width=480*5,
431     height = 240*5, res=120)
432 ggpairs(vehicle[,1:ncol(vehicle)-1], mapping=ggplot2::
433     aes(colour='red', alpha=0.05),

```

```

427         upper = list(corrMethod = "kendall"))
428 dev.off()
429
430 # Dibujar un gráfico de correlaciones basados en
431 # Spearman
432 corr <- cor(vehicle[,1:ncol(vehicle)-1], method='kendall')
433
434 pdf('Correlograma_spearman_vehicle.pdf', width=14,
435      height = 7)
436 par(mfrow=c(1,2))
437 corrrplot(corr, method='ellipse', type='lower')
438 corrrplot(corr, method='number', type='lower', number.cex
439             = 0.80)
440 dev.off()

```

E. Script5.R

Este *script* fue usado para la elaboración de modelos de clasificación sobre el *dataset* *vehicle*.

```

1 # Librerías importadas
2 library('caret')
3 library('class')
4 library('MASS')
5 library('dplyr')
6 library('MVTests')
7
8 # Función para leer el dataset vehicle
9 read.vehicle.dataset <- function(filename) {
10   # Cargar datos
11   dat <- read.table(filename, comment.char="@", sep=',')
12
13   # Asignar nombres
14   names(dat) <- c('Compactness', 'Circularity',
15                   'Distance_circularity',
16                   'Radius_ratio', 'Praxis_aspect_ratio',
17                   'Max_length_aspect_ratio', 'Scatter_
18                   ratio',
19                   'Elongatedness', 'Praxis_rectangular',
20                   'Length_rectangular',
21                   'Major_variance', 'Minor_variance',
22                   'Gyration_radius',
23                   'Major_skewness', 'Minor_skewness',
24                   'Minor_kurtosis',
25                   'Major_kurtosis', 'Hollows_ratio',
26                   'Class')

```

```

21     return(dat)
22 }
23 }

24 # Cargar el dataset vehicle
25 vehicle <- read.vehicle.dataset('./Datasets
26             Clasificacion/vehicle/vehicle.dat')
27 fold_basename <- './Datasets Clasificacion/vehicle/
28             vehicle'

29 # Modelos basados en k-NN
30
31 # Evaluar los modelos mediante validación cruzada 10-
32     fold
33 run_knn_k_fold_vehicle <- function(i, x, k, standarizate
34 =NULL, tt = "test") {
35     # Cargar conjuntos de entrenamiento
36     file <- paste(x, "-10-", i, "tra.dat", sep="");
37     x_tra <- read.csv(file, comment.char="@")
38
39     # Cargar conjuntos de test
40     file <- paste(x, "-10-", i, "tst.dat", sep="");
41     x_tst <- read.csv(file, comment.char="@")
42
43     names(x_tra) <- c('Compactness', 'Circularity',
44             'Distance_circularity',
45                     'Radius_ratio', 'Praxis_aspect_ratio
46             ,
47                     'Max_length_aspect_ratio', 'Scatter_
48             ratio',
49                     'Elongatedness', 'Praxis_rectangular
50             , 'Length_rectangular',
51                     'Major_variance', 'Minor_variance',
52             'Gyration_radius',
53                     'Major_skewness', 'Minor_skewness',
54             'Minor_kurtosis',
55                     'Major_kurtosis', 'Hollows_ratio',
56             'Class')
57
58     names(x_tst) <- names(x_tra)
59
60     # Estandarizar las variables proporcionadas
61     for (i in standarizate) {
62         x_tra[i] <- scale(x_tra[i])
63         x_tst[i] <- scale(x_tst[i])
64     }
65
66     print(summary(x_tra))
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
589

```

```

59 if (tt == "train") { test <- x_tra }
60 else { test <- x_tst }

61
62 # Entrenar el modelo sobre el conjunto de train
63 model.eval <- knn(train=x_tra[,-ncol(x_tra)], test=x_
64   tst[,-ncol(x_tst)],
65           cl=x_tra[,ncol(x_tra)], k=k)

66 # Evaluación del CCR sobre test
67 postResample(model.eval, test$Class)[1]
68 }

69
70 # Modelo con k=7
71 knn.model.fit1 <- knn(train=vehicle[,-ncol(vehicle)],
72   test=vehicle[,-ncol(vehicle)],
73           cl=vehicle[,ncol(vehicle)], k=7)

74 # Evaluar error CCR train
75 cat('Accuracy sobre el conjunto de train: ',postResample
76   (knn.model.fit1,
77
78   vehicle$Class)[1],
79   fill=T)
80
81
82
83 # Modelo con k=7, con atributos escalados
84 vehicle.rescaled <- vehicle
85 vehicle.rescaled[,-ncol(vehicle)] <- scale(vehicle.
86   rescaled[,-ncol(vehicle)])
87
88 knn.model.fit2 <- knn(train=vehicle.rescaled[,-ncol(
89   vehicle.rescaled)],
90           test=vehicle.rescaled[,-ncol(
91   vehicle.rescaled)],
92           cl=vehicle.rescaled[,ncol(vehicle.
93   rescaled)], k=7)

94 # Evaluar error CCR train
95 cat('Accuracy sobre el conjunto de train: ',postResample
96   (knn.model.fit2,
97
98   vehicle.
99   rescaled$Class)[1],
100  fill=T)

```

```

95 cat('Accuracy sobre 10-fold: ',mean(sapply(1:10,run_knn_
96   k_fold_vehicle,
97   , 7,
98   , fold_basename
99   , colnames(
100   vehicle)[1:ncol(vehicle)-1])),
101   fill=T)
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
```

```

132 # Modelos LDA
133
134 # Analizar las varianzas por clase
135 vehicle.bus <- vehicle %>% filter(Class==' bus ')
136 vehicle.opel <- vehicle %>% filter(Class==' opel')
137 vehicle.saab <- vehicle %>% filter(Class==' saab')
138 vehicle.van <- vehicle %>% filter(Class==' van ')
139
140 var.class <- rbind(apply(vehicle.bus[,-ncol(vehicle.bus)]
141   ], FUN=var, MARGIN = 2),
142     apply(vehicle.opel[,-ncol(vehicle.
143       opel)], FUN=var, MARGIN = 2),
144     apply(vehicle.saab[,-ncol(vehicle.
145       saab)], FUN=var, MARGIN = 2),
146     apply(vehicle.van[,-ncol(vehicle.van)
147       ], FUN=var, MARGIN = 2))
148
149 rownames(var.class) <- c('bus', 'opel', 'saab', 'van')
150 var.class
151
152 # Analizar las equidad en las covarianzas con el test de
153 # BoxM
154 test.boxm <- BoxM(vehicle[,-ncol(vehicle)], group=
155   vehicle$Class)
156 cat('p-value asociado al test BoxM: ',test.boxm$p.value)
157
158 # Función para validación cruzada
159 run_lda_k_fold_vehicle <- function(i, x, model, tt =
160   "test") {
161   # Cargar conjuntos de entrenamiento
162   file <- paste(x, "-10-", i, "tra.dat", sep="");
163   x_tra <- read.vehicle.dataset(file)
164
165   # Cargar conjuntos de test
166   file <- paste(x, "-10-", i, "tst.dat", sep="");
167   x_tst <- read.vehicle.dataset(file)
168
169   if (tt == "train") { test <- x_tra }
170   else { test <- x_tst }
171
172   # Entrenar el modelo sobre el conjunto de train
173   form <- terms(model)
174   model.eval <- lda(formula=form, data=x_tra)
175
176   # Evaluación del CCR sobre test
177   pred <- predict(model.eval, test)
178   postResample(pred$class, test$Class)[1]
179 }

```

```

174 # Class ~ .
175 lda.model.fit1 <- lda(Class~.-Scatter_ratio-Praxis_
176   rectangular-Length_rectangular-
177   Minor_variance-Gyration_radius,
178   data=vehicle)
179 lda.model.fit1
178
179 lda.model.fit1.pred <- predict(lda.model.fit1, vehicle)
180 cat('CCR medido sobre entrenamiento:', postResample(lda.
181   model.fit1.pred$class,
181
182   vehicle$Class)[1], fill=T)
182 cat('Accuracy sobre 10-fold: ', mean(sapply(1:10, run_lda_
183   k_fold_vehicle,
183
184   , lda.model.fit1)), fill=T)
184 # Class ~ .
185 lda.model.fit2 <- lda(Class~, data=vehicle)
186 lda.model.fit2
187
188 lda.model.fit2.pred <- predict(lda.model.fit2, vehicle)
189 cat('CCR medido sobre entrenamiento:', postResample(lda.
190   model.fit2.pred$class,
190
191   vehicle$Class)[1], fill=T)
191 cat('Accuracy sobre 10-fold: ', mean(sapply(1:10, run_lda_
192   k_fold_vehicle,
192
193   , lda.model.fit2)), fill=T)
193
194 # Modelos QDA
195
196 # Analizar las correlaciones entre las variables para
197   cada clase
197 corr.bus <- cor(vehicle.bus[,-ncol(vehicle.bus)], method
198   = 'kendall')
198 corr.opel <- cor(vehicle.opel[,-ncol(vehicle.opel)], method
199   = 'kendall')
199 corr.saab <- cor(vehicle.saab[,-ncol(vehicle.saab)], method
200   = 'kendall')
200 corr.van <- cor(vehicle.van[,-ncol(vehicle.van)], method
201   = 'kendall')
201
202 pdf('Correlograma_kendall_vehicle_clases.pdf', width=14,
203   height = 15)
203 par(mfrow=c(2,2))
204 corrrplot(corr.bus, method='color', type='lower', number.
204   cex = 0.8,

```

```

205         title='bus', addCoef.col='gray', mar=c(0,0,1,0)
206     )
207 corrrplot(corr.opel, method='color', type='lower', number
208 .cex = 0.80,
209         title='opel', addCoef.col='gray', mar=c
210 (0,0,1,0))
211 corrrplot(corr.saab, method='color', type='lower', number
212 .cex = 0.80,
213         title='saab', addCoef.col='gray', mar=c
214 (0,0,1,0))
215 corrrplot(corr.van, method='color', type='lower', number.
216 cex = 0.80,
217         title='van', addCoef.col='gray', mar=c(0,0,1,0))
218 dev.off()
219
220 # Elaborar una función de validación para qda
221 run_qda_k_fold_vehicle <- function(i, x, model, tt =
222   "test") {
223   # Cargar conjuntos de entrenamiento
224   file <- paste(x, "-10-", i, "tra.dat", sep="");
225   x_tra <- read.vehicle.dataset(file)
226
227   # Cargar conjuntos de test
228   file <- paste(x, "-10-", i, "tst.dat", sep="");
229   x_tst <- read.vehicle.dataset(file)
230
231   if (tt == "train") { test <- x_tra }
232   else { test <- x_tst }
233
234   # Entrenar el modelo sobre el conjunto de train
235   form <- terms(model)
236   model.eval <- qda(formula=form, data=x_tra)
237
238   # Evaluación del CCR sobre test
239   pred <- predict(model.eval, test)
240   postResample(pred$class, test$Class)[1]
241 }
242
243 # Elaborar un modelo basado en QDA
244 # Class ~ .
245 qda.model.fit1 <- qda(Class~, data=vehicle)
246 qda.model.fit1
247
248 qda.model.fit1.pred <- predict(qda.model.fit1, vehicle)
249 cat('CCR medido sobre entrenamiento:', postResample(qda.
250     model.fit1.pred$class,
251
252     vehicle$Class)[1], fill=T)

```

```
244 cat('Accuracy sobre 10-fold: ',mean(sapply(1:10,run_qda_
k_fold_vehicle,
245 , qda.model.fit1)),fill=T)
          fold_basename
```