



# UNIVERSIDAD DE GRANADA

INTRODUCCIÓN A LA CIENCIA DE DATOS  
MÁSTER CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

---

## TRABAJO TEÓRICO/PRÁCTICO

ANÁLISIS DE DATOS, REGRESIÓN Y CLASIFICACIÓN

---

### Autor

Ignacio Vellido Expósito  
ignaciove@correo.ugr.es



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE  
TELECOMUNICACIÓN

CURSO 2020-2021

# Índice

<b>1. Regresión: Análisis Estadístico de Datos</b>	<b>2</b>
1.1. Introducción . . . . .	2
1.2. Análisis Estadístico de Datos . . . . .	3
1.2.1. Análisis univariable . . . . .	3
1.2.2. Análisis sobre las distribuciones . . . . .	13
1.2.3. Transformaciones . . . . .	17
1.2.4. Outliers . . . . .	18
1.2.5. Análisis de correlación . . . . .	18
1.2.6. Tratamiento de variables . . . . .	25
1.2.7. Ordenaciones . . . . .	25
1.2.8. Resolución de hipótesis . . . . .	25
1.3. Conclusiones . . . . .	31
<b>2. Técnicas de Regresión</b>	<b>32</b>
2.1. Ajustes de regresión lineal univariantes . . . . .	34
2.2. Ajustes de regresión lineal multivariable . . . . .	37
2.3. Inserción de interacciones . . . . .	39
2.4. Ajustes de regresión no lineal . . . . .	42
2.5. Ajustes con KNN . . . . .	50
2.6. Comparativa de los ajustes anteriores con cross-validation . . . . .	55
2.7. Comparativa de tests . . . . .	55
<b>3. Clasificación: Análisis Estadístico de Datos</b>	<b>57</b>
<b>4. Técnicas de Clasificación</b>	<b>58</b>
<b>Referencias</b>	<b>59</b>

# 1. Regresión: Análisis Estadístico de Datos

## 1.1. Introducción

Para el problema de regresión hacemos uso del dataset **autoMPG6** [1], donde se codifica el consumo de gasolina de distintos coches (en millas por galón, Mpg) en base a las siguientes características:

1. **Displacement**: Indica la cilindrada del coche, la suma del volumen útil de los cilindros del motor, medido en pulgadas cúbicas.
2. **Horse\_power**: Mide la potencia del coche.
3. **Weight**: Peso en libras.
4. **Acceleration**: Aceleración del coche de 0 a 60 millas por hora, medido en segundos.
5. **Model\_year**: Indica las dos últimas cifras del año de producción.

El objetivo es poder predecir con estos cinco atributos el consumo de Mpg de un nuevo coche:

6. **Mpg**: Millas-por-galón, indica la cantidad de galones ( $1\text{G} \approx 3,78\text{L}$ ) de fuel que consume un vehículo al recorrer una milla ( $1\text{m} \approx 1,6\text{km}$ ).

El dataset contiene 392 instancias codificando esta información.

La descripción del problema nos da alguna información adicional sobre las variables:

1. **Displacement**: Variable numérica continua, contamos con valores reales en el rango  $[68.0, 455.0]$ .
2. **Horse\_power**: Variable numérica continua, contamos con valores enteros en el rango  $[46, 230]$ .
3. **Weight**: Variable numérica continua, contamos con valores enteros en el rango  $[1613, 5140]$ .
4. **Acceleration**: Variable numérica continua, contamos con valores reales en el rango  $[8.0, 24.8]$ .
5. **Model\_year**: Variable numérica discreta, contamos con valores enteros en el rango  $[70, 82]$ .
6. **Mpg**: Variable numérica continua, contamos con valores reales en el rango  $[9.0, 46.6]$ .

### Hipótesis de partida

- **H.1**: Horse\_power puede influir en Mpg: A más potencia, más consumo.
- **H.2**: Weight debe influir en Mpg: Un coche más pesado debería consumir más.
- **H.3**: Debería haber correlación entre displacement (cilindrada) con horse y acceleration

- **H.4:** Horse y acceleration podrían estar relacionadas
- **H.5:** Viendo que contamos con un rango pequeño de años, no debería haber un cambio significativo de prestaciones entre años
- **H.6:** Pero debería existir una tendencia de mejora de prestaciones con los años, incluyendo aumento de Displacement, Horse\_power y Acceleration.
- **H.7:** Model\_year podría no mostrar relación con Mpg: Pese al paso de los años si contamos con diferentes tipos de vehículos (todoterrenos, familiares, deportivos...) podría haber un consumo dispar. (Si existiera tendencia, viendo que los años son de las últimas décadas del siglo XX, podría ir el consumo hacia abajo)
- **H.8:** Esta última hipótesis se puede aplicar al resto de variables, indicándonos que Model\_year no debería tener relevancia para este problema de regresión.
- **H.9:** Horse\_power podría depender de las variables Displacement y Weight

## 1.2. Análisis Estadístico de Datos

Antes de comenzar a analizar las variables nos planteamos una cuestión: ¿Debemos considerar Model\_year como una variable numérica o como un factor categórico? Aunque por la hipótesis H.7 podríamos acabar no eligiendo la variable para el problema, es necesario preguntarnos por esto antes de comenzar.

Sabemos que las observaciones para esta variable cuenta con valores entre 72 y 82, por lo que tenemos información exacta del año (en comparación, por ejemplo, con agrupaciones mayores como la década o el siglo). El hecho de tratarla como categórica o cuantitativa depende mucho del problema. En este caso, tenemos interés en cuestionarnos por valores entre años, por ejemplo, el consumo entre los años 75 y 76.

Por tanto, de cara al problema de regresión que nos atañe, tendríamos dos opciones:

- Mantenerlo como categórico y generar variables dummy (valores 0-1 para indicar si la instancia es de ese año). Suponiendo que tenemos al menos una instancia de cada año, esto nos generaría 12 variables nuevas.
- Mantenerlo como numérico, pero teniendo cuidado de cómo interpretar el año.

Proseguimos con tanto dejando Model\_year como variable numérica.

### 1.2.1. Análisis univariable

La cabecera de nuestro dataset tiene esta forma:

Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
91	70	1955	20.5	71	26.0
232	100	2789	15.0	73	18.0
350	145	4055	12.0	76	13.0
318	140	4080	13.7	78	17.5
113	95	2372	15.0	70	24.0
97	60	1834	19.0	71	27.0

Y con la siguiente información estadística:

Displacement	Horse_power	Weight	Acceleration	Model_year
Min. : 68.0	Min. : 46.0	Min. : 1613	Min. : 8.00	Min. : 70.00
1st Qu.: 105.0	1st Qu.: 75.0	1st Qu.: 2225	1st Qu.: 13.78	1st Qu.: 73.00
Median : 151.0	Median : 93.5	Median : 2804	Median : 15.50	Median : 76.00
Mean : 194.4	Mean : 104.5	Mean : 2978	Mean : 15.54	Mean : 75.98
3rd Qu.: 275.8	3rd Qu.: 126.0	3rd Qu.: 3615	3rd Qu.: 17.02	3rd Qu.: 79.00
Max. : 455.0	Max. : 230.0	Max. : 5140	Max. : 24.80	Max. : 82.00

Mpg
Min. : 9.00
1st Qu.: 17.00
Median : 22.75
Mean : 23.45
3rd Qu.: 29.00
Max. : 46.60

El dataset **no** cuenta con **valores repetidos** ni **missing values**.

Mostramos scatterplots univariates:

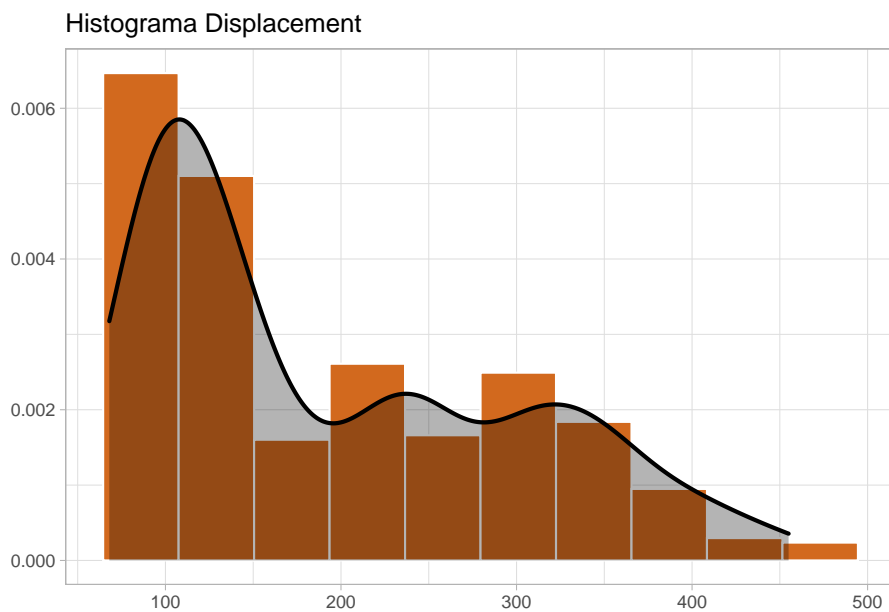


Figura 1

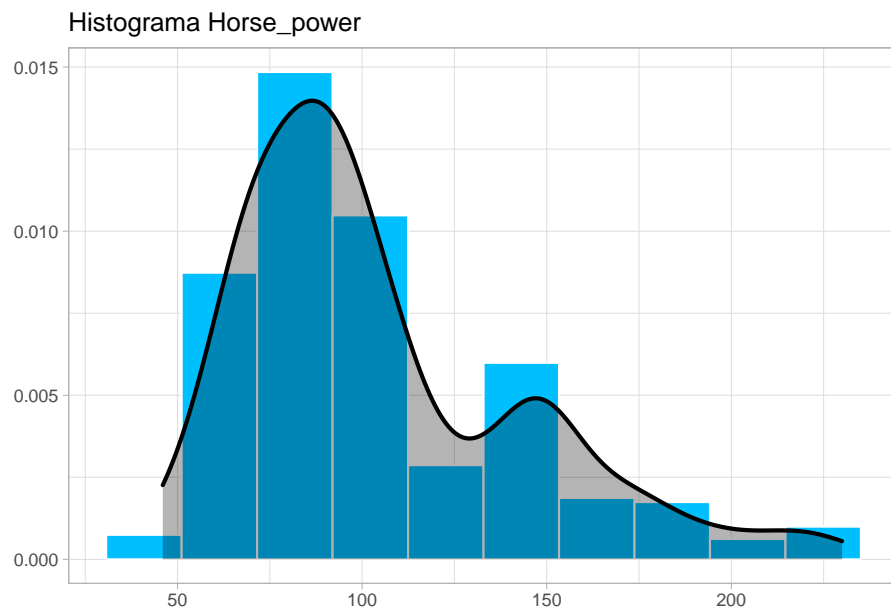


Figura 2

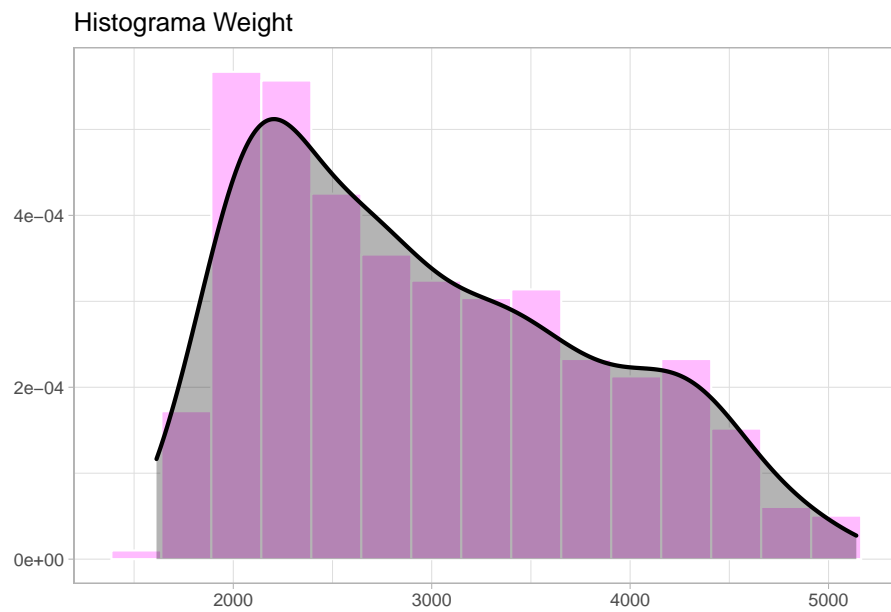


Figura 3

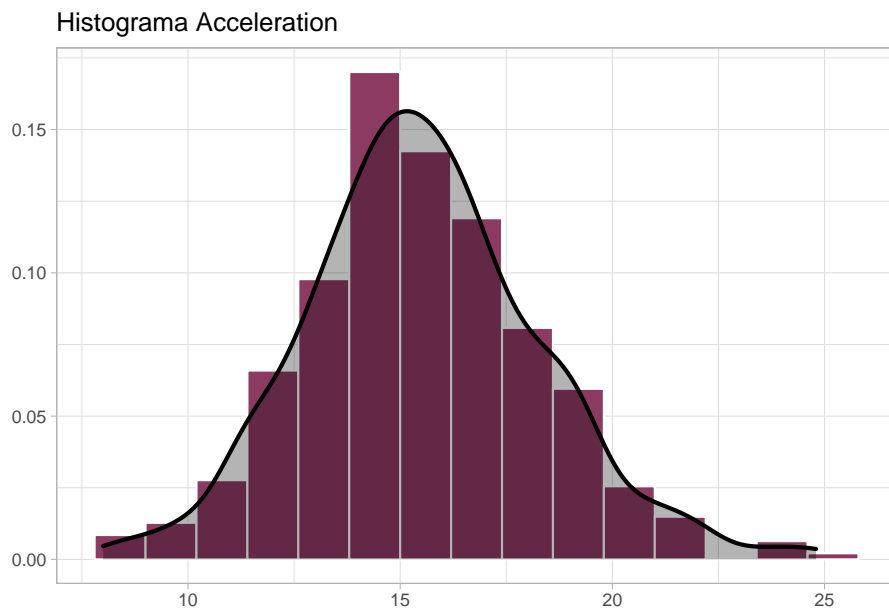


Figura 4

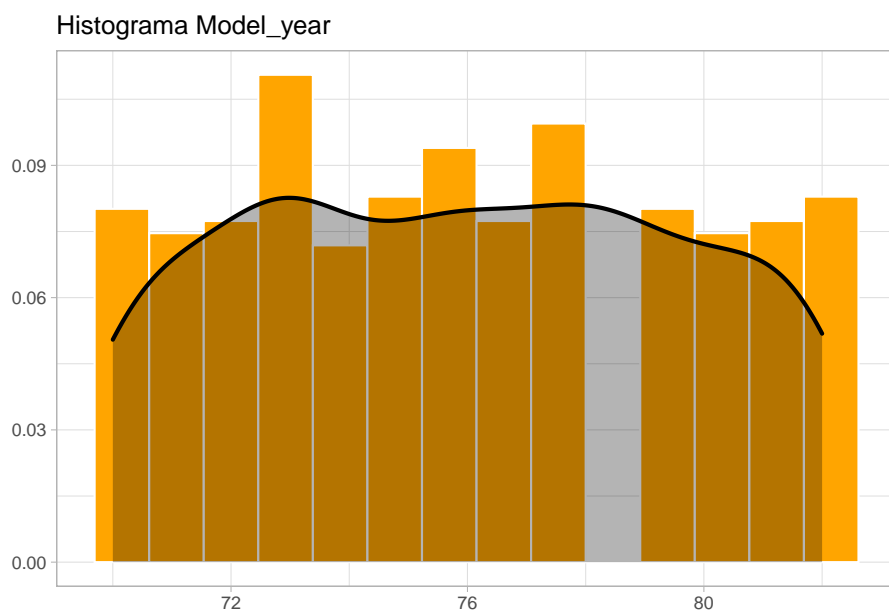


Figura 5

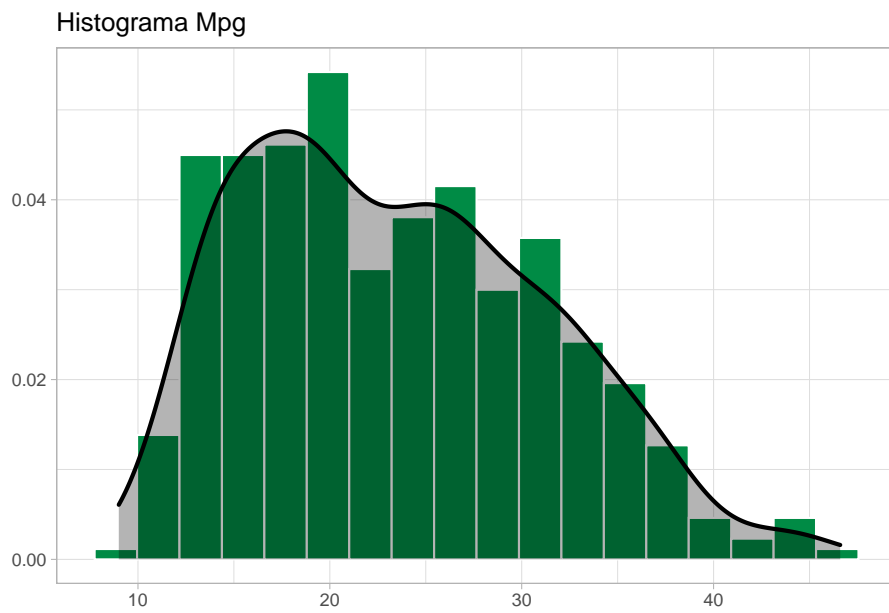


Figura 6

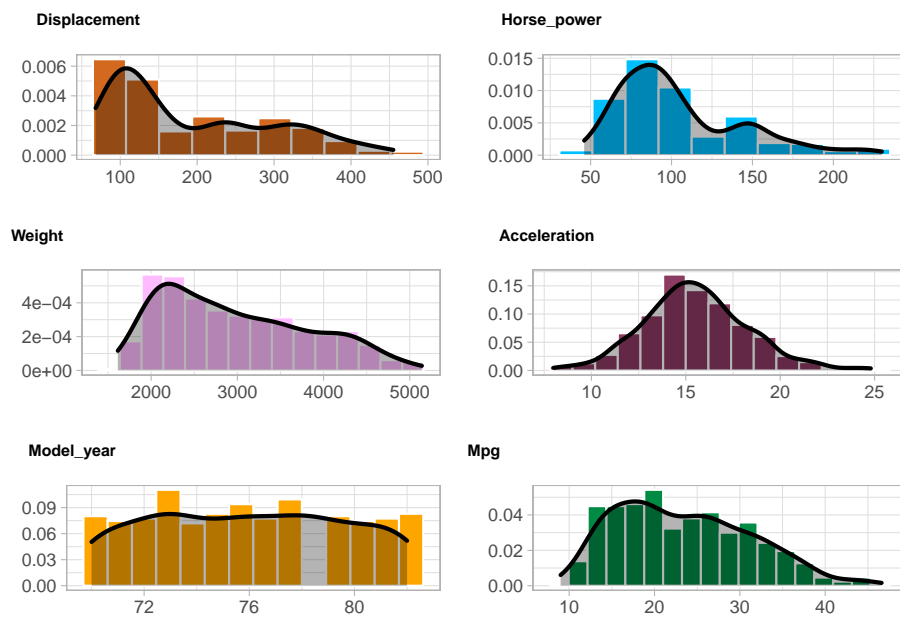


Figura 7

Y boxplots sobre las distribuciones de los datos:



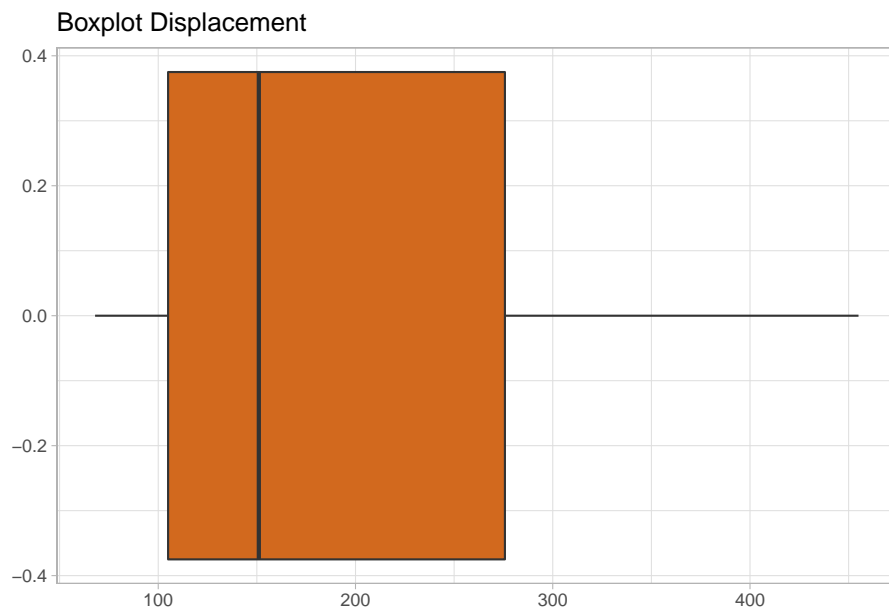


Figura 8

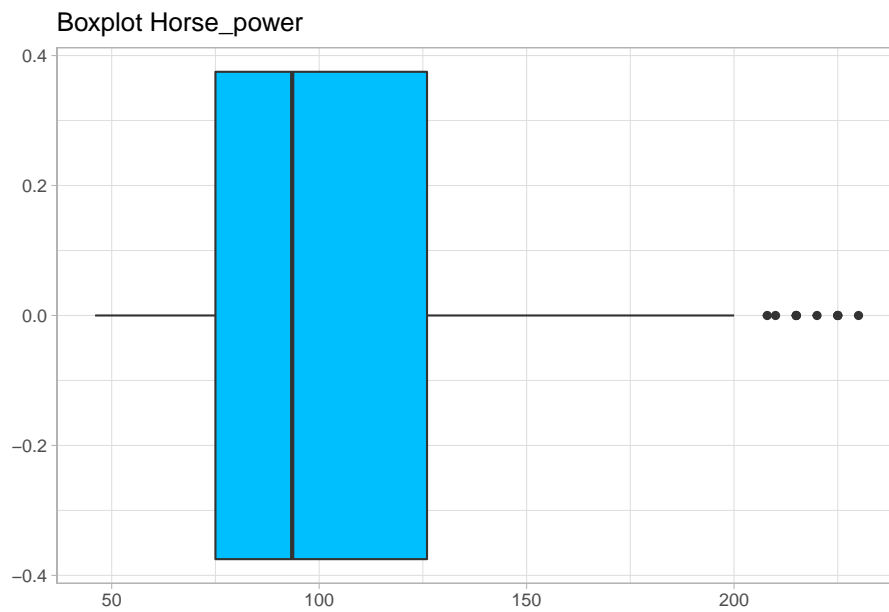


Figura 9

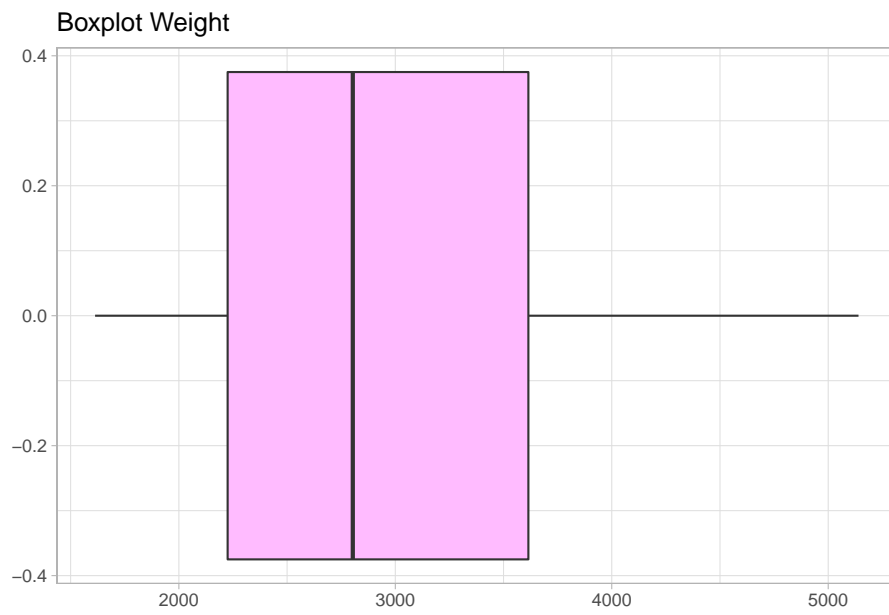


Figura 10

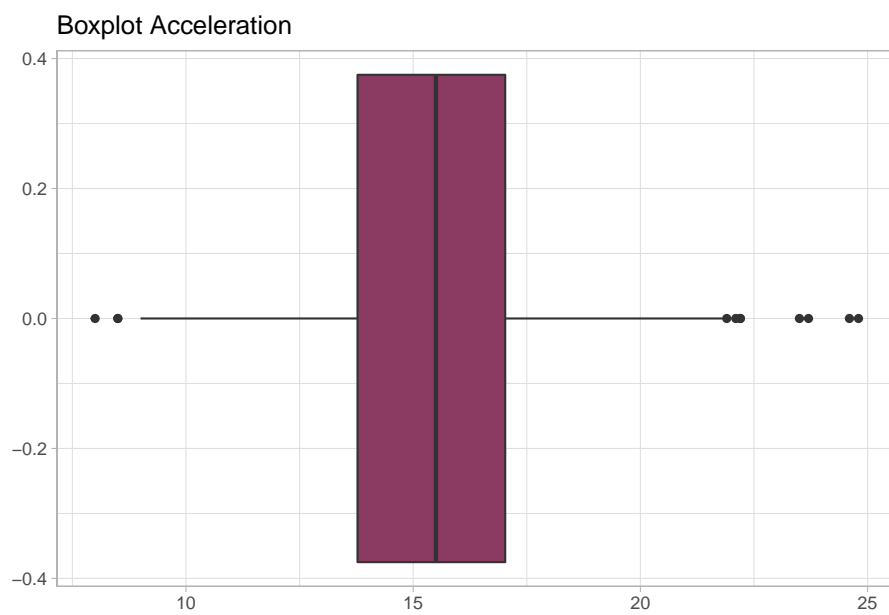


Figura 11

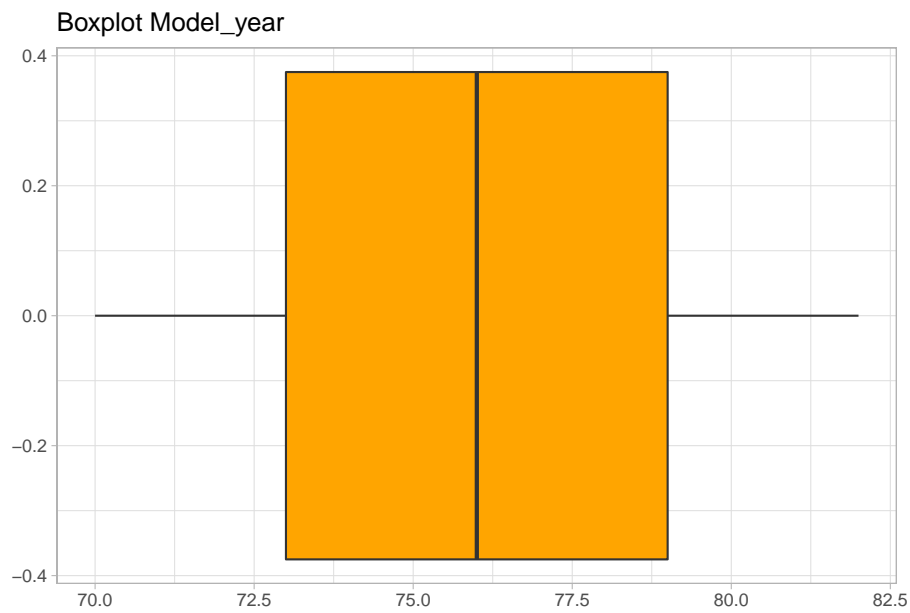


Figura 12

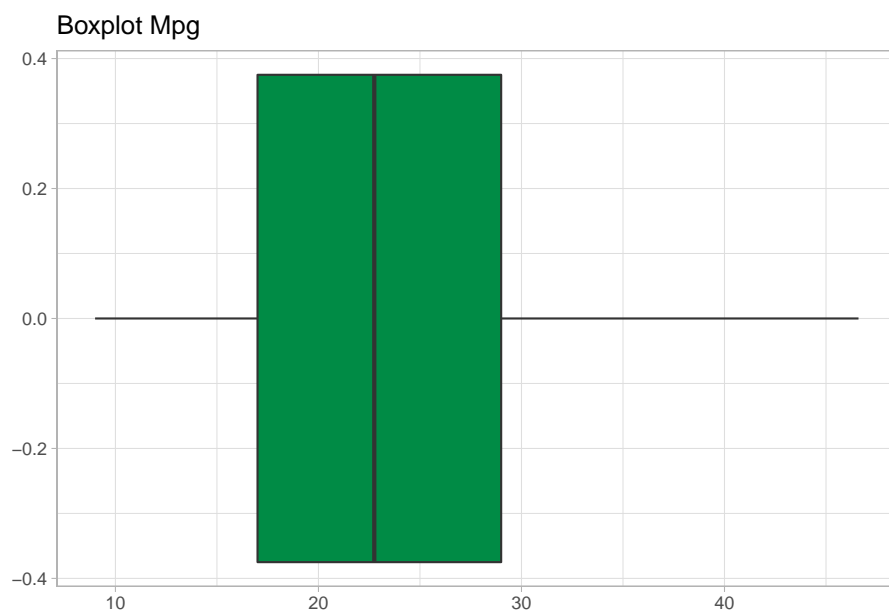


Figura 13

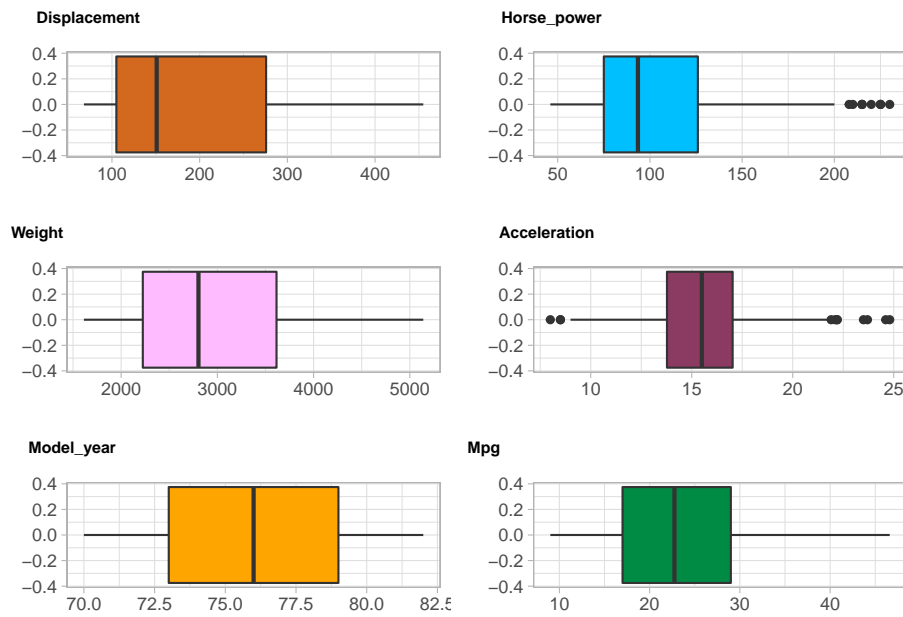


Figura 14

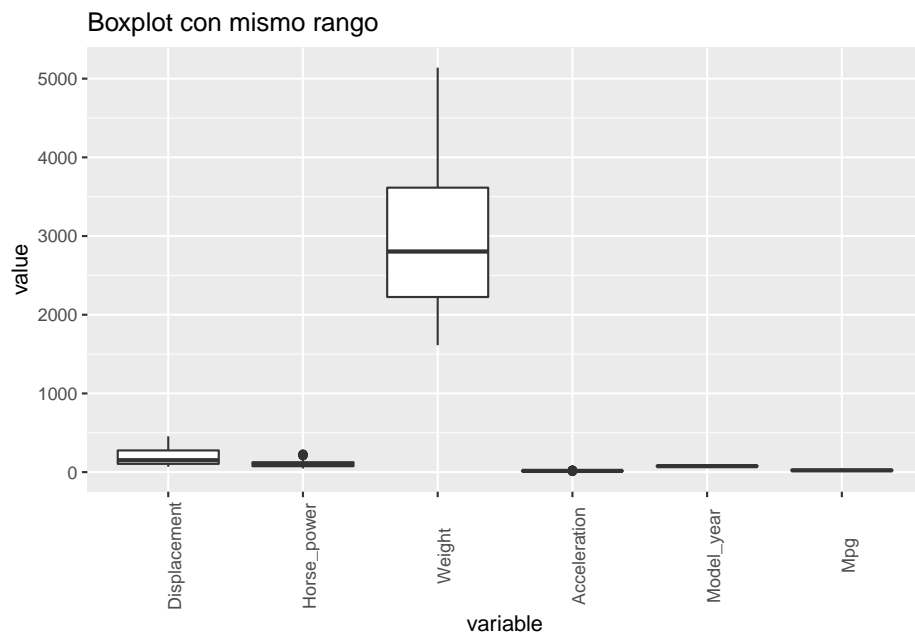


Figura 15

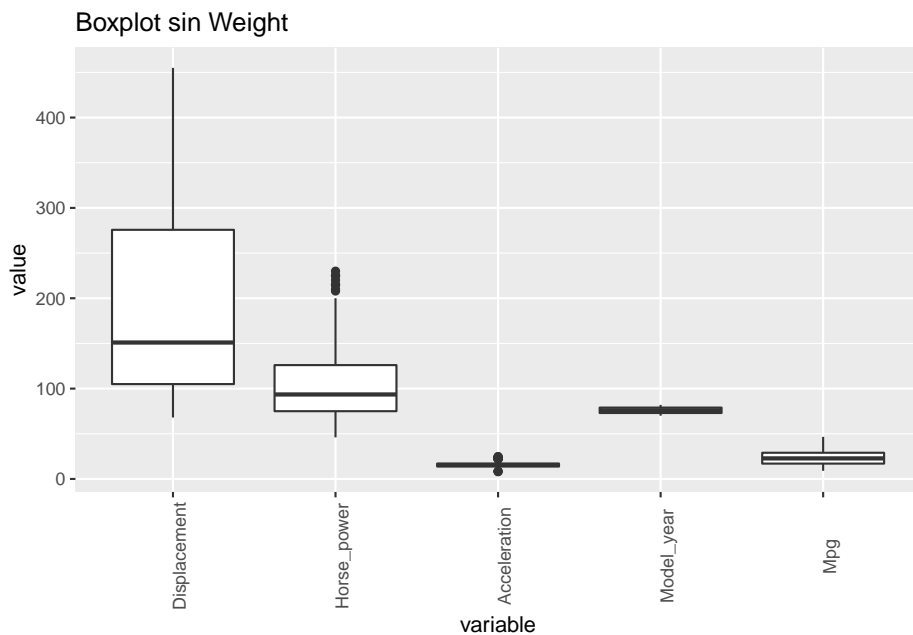


Figura 16

Ya la descripción del problema nos lo decía, los rangos en los que se distribuyen los datos son muy diferentes dependiendo de la variable. Se pueden estandarizar los datos para solucionar este problema, aunque para regresión lineal no es necesario (sí lo es para KNN). Podemos comparar los rangos intercuantiles si estandarizamos antes el dataset

Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1.631723	1.324980	1.635856	1.178021	1.628781	1.537475

También podemos ver la distancia entre mínimos y máximos

Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
3.698253	4.780318	4.152330	6.089463	3.257562	4.817420

**Displacement** La cilindrada vemos con una desviación grande y una gran concentración en los valores inferiores. Desviado a la izquierda, no parece seguir una distribución normal. Existe una alta concentración en torno al valor 125, muy por encima del recuento que alcanzan el resto de valores

**Horse\_power** Similar a Displacement pero cuenta con una mayor dispersión y algunos valores muy altos. A día de hoy los coches suelen rondar los 120 en turismos y los 200 en SUVs. Aquí contamos con predominancia en el rango aproximado [70, 125] con algunas instancias por encima de los 200. Desviado a la izquierda, no parece seguir una distribución normal.

**Weight** Una distribución más achatada que las anteriores, también ladeada hacia la izquierda. Un rango mayor

**Acceleration** Valores altamentes concentrados pero en general con un rango alto. Su forma se asemeja a una distribución normal.

**Model\_year** Aunque no se vea bien en las gráficas, contamos con valores de todos los años, más o menos equitativamente

Años: 70 71 72 73 74 75 76 77 78 79 80 81 82  
 Conteo: 29 27 28 40 26 30 34 28 36 29 27 28 30

### 1.2.2. Análisis sobre las distribuciones

Hemos comentado antes que no apreciamos semejanzas con una distribución normal en algunas de las variables, lo comprobamos con un test estadístico (Shapiro-Wilk test):

vars	statistic	p_value	sample
Displacement	0.8818359	0.0000000	392
Horse_power	0.9040975	0.0000000	392
Weight	0.9414661	0.0000000	392
Acceleration	0.9918671	0.0305289	392
Model_year	0.9469666	0.0000000	392
Mpg	0.9671696	0.0000001	392

El test de Shapiro nos asegura con bastante certeza que ninguna variable sigue una distribución normal, aunque en menor grado en Acceleration (sigue siendo al 97% de confianza).

Para los modelos de regresión que vamos a usar aún así la normalidad de los datos no es necesaria. Se muestra aquí como no hay que dejarse engañar por los gráficos, puesto

que Acceleration parecía seguirla. El p-value de Acceleration está muy cerca del umbral (0.03 vs 0.05). Es bastante probable de que la parte central derecha de la distribución sea la causante de no asegurar la normalidad.

Mostramos con gráficos Q-Q cómo se separan las distribuciones de su supuesta normal:

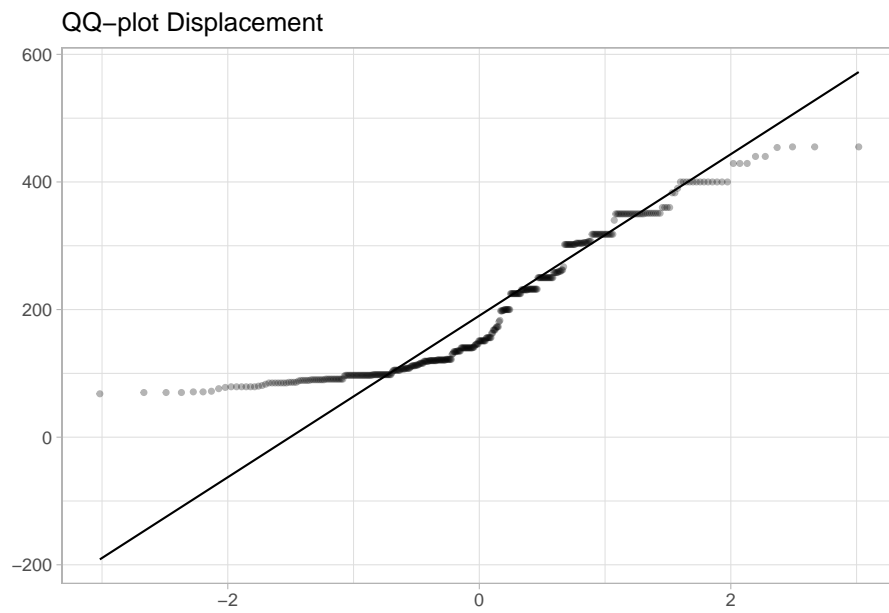


Figura 17

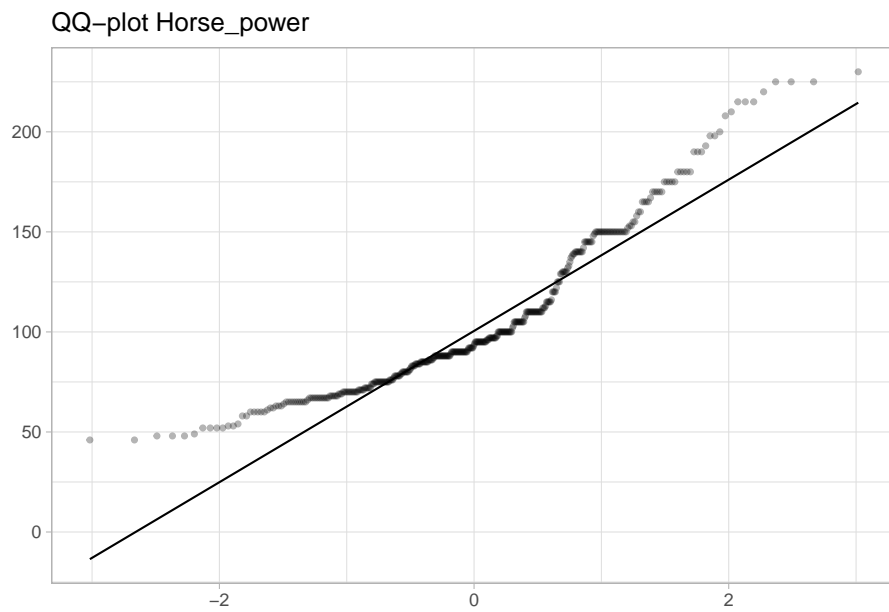


Figura 18

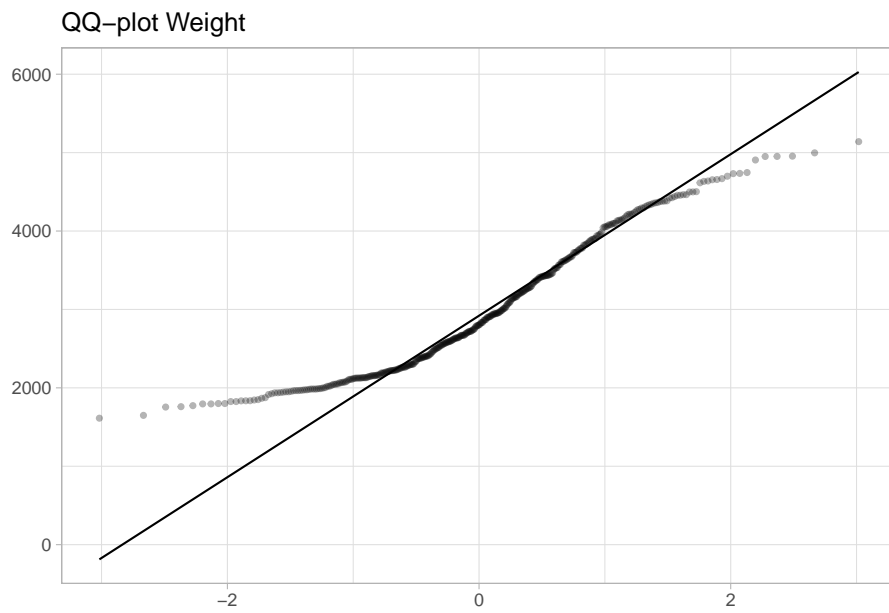


Figura 19

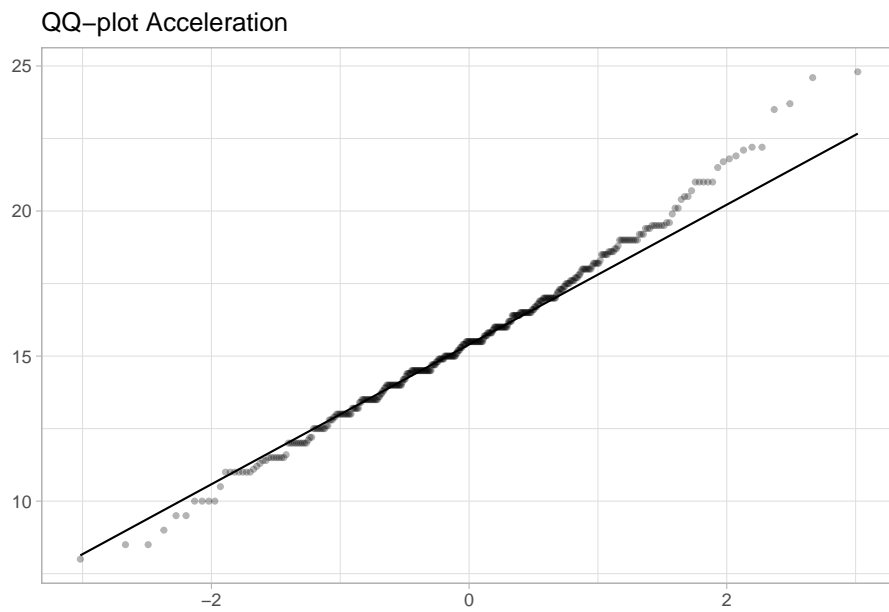


Figura 20



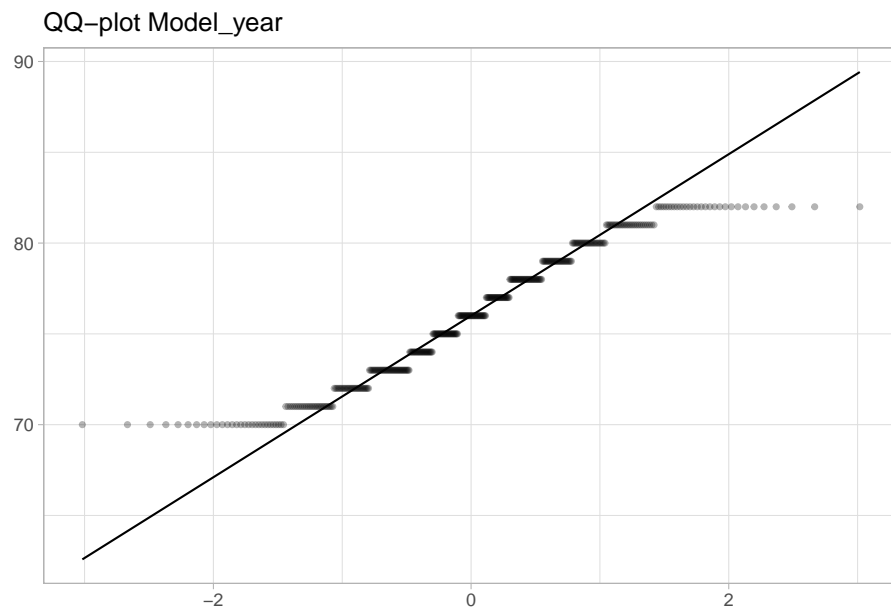


Figura 21

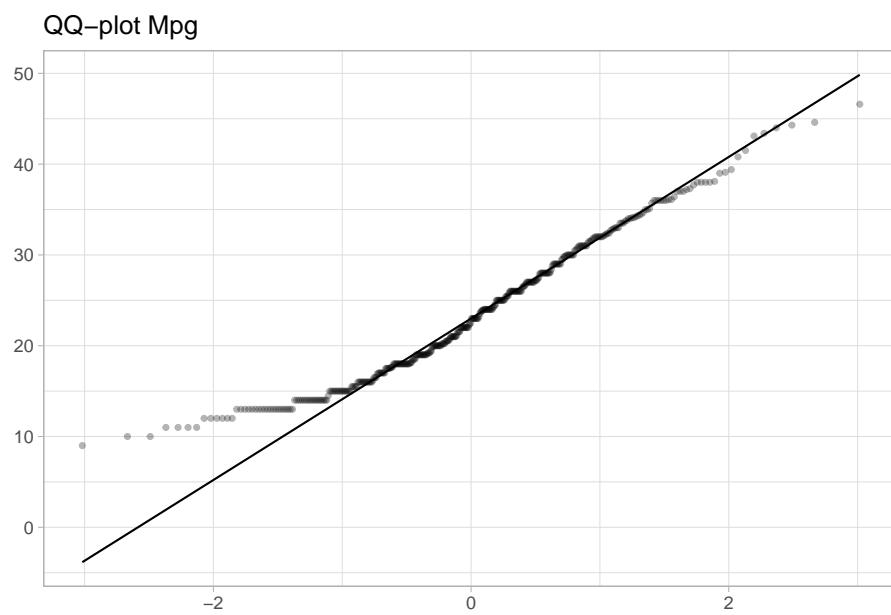


Figura 22

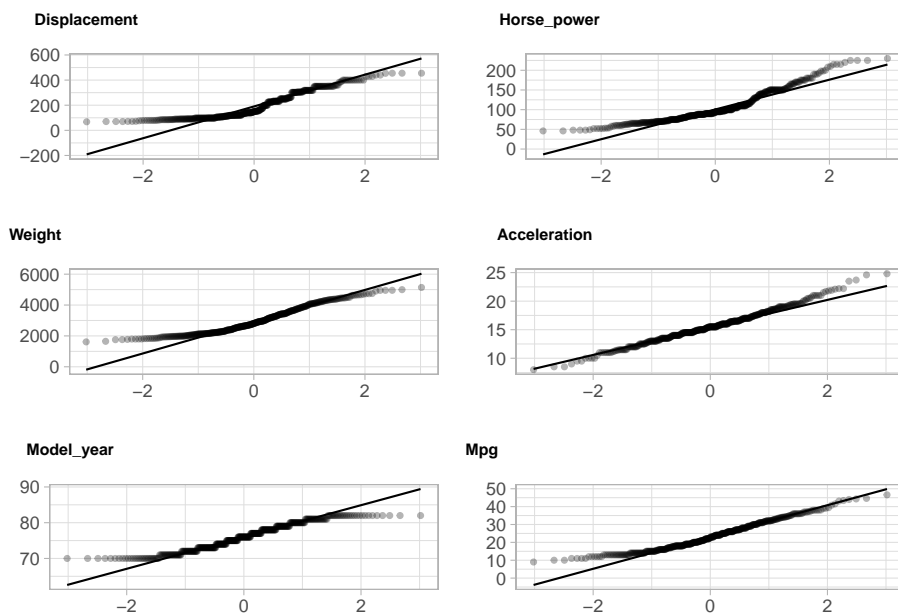


Figura 23

Estos gráficos Q-Q nos muestran más claramente que las variables no siguen distribuciones normales. La distribución de Acceleration es la que más se asemeja y eso lo vemos en el estadístico de Shapiro, pero en la cola superior existe una diferencia significativa que hace que el test rechace.

**Skewness** Las variables con skewness que tenemos son:

```
"Displacement" "Horse_power" "Weight"
```

Concretamente, sus valores son:

```
Displacement: [1] 0.6989813
Horse_power: [1] 1.083161
Weight: [1] 0.5175953
```

Adicionalmente, se muestra:

```
Mpg: [1] 0.4553414
```

Sobre la skewness, tal y como se había visto en las gráficas, algunas de las variables la tienen, en los 3 casos positivas (hacia la izquierda).

Los plots nos han dado idea de que Mpg tiene cierta skewness, pero cae por debajo del umbral de 0.5.

### 1.2.3. Transformaciones

En general no consideramos ninguna transformación necesaria para el dataset con el que contamos. No vemos necesario crear variables nuevas a partir de las vistas, puesto que por el conocimiento que tenemos del problema parece que las variables son coherentes.

Las transformaciones necesarias para pasar a una distribución normal dependen de la variable en cuestión. Primero deberíamos averiguar que tipo de distribución siguen. Aunque, tal y como se ha comentado anteriormente, los métodos utilizados para regresión (regresión lineal y KNN) no asumen ninguna forma para la distribución de los datos, por lo que no es necesario aplicar nada.

Adicionalmente, aunque para regresión lineal no es absolutamente necesario, podemos estandarizar los datos a media 0 y dev 1, facilitando un poco los cálculos. La inferencia estadística de la regresión no variaría aplicando esta transformación, pero deberíamos tener cuidado a la hora de interpretar los resultados de la regresión para no confundirnos.

#### 1.2.4. Outliers

Como hemos visto anteriormente en los boxplots, las únicas variables con valores muy alejados del centro de la distribución son Acceleration y Horse\_power.

Por el significado del problema, probablemente estos posibles outliers correspondan a coches de alta gama o potentes en la época. Esto tampoco lo podemos asegurar puesto que contamos con pocas características, pero se considera un razonamiento coherente. Además, puesto que los valores caen dentro de los rangos posibles para coches de la época, podemos descartar que sean errores de medida.

Deberíamos decidir si mantener o no estas instancias. Como en nuestro caso se nos ha pedido predecir el consumo Mpg, sin darnos consideraciones sobre los tipos/gamas de coches a los que se enfoca, proseguimos dejándo estas filas.

#### 1.2.5. Análisis de correlación

Tenemos que tener en cuenta que las variables no siguen distribuciones normales. Aunque el coeficiente de Pearson no asume normalidad (si asume varianza y covarianza finitas), podemos además usar el coeficiente de Kendall para los cálculos. Independientemente del método usado vamos a obtener las mismas correlaciones en este dataset, solo varía el valor de fuerza con la que se dan.

Para regresión la correlación en los datos no es preocupante. Al contrario, podría haber información (poca, pero alguna cantidad) que se aporte y nos ayude en el problema. En el peor de los casos, la propia metodología de selección de variables en el modelo multivariable nos ayudará a descartar aquellas variables que no sean necesarias como regresor.

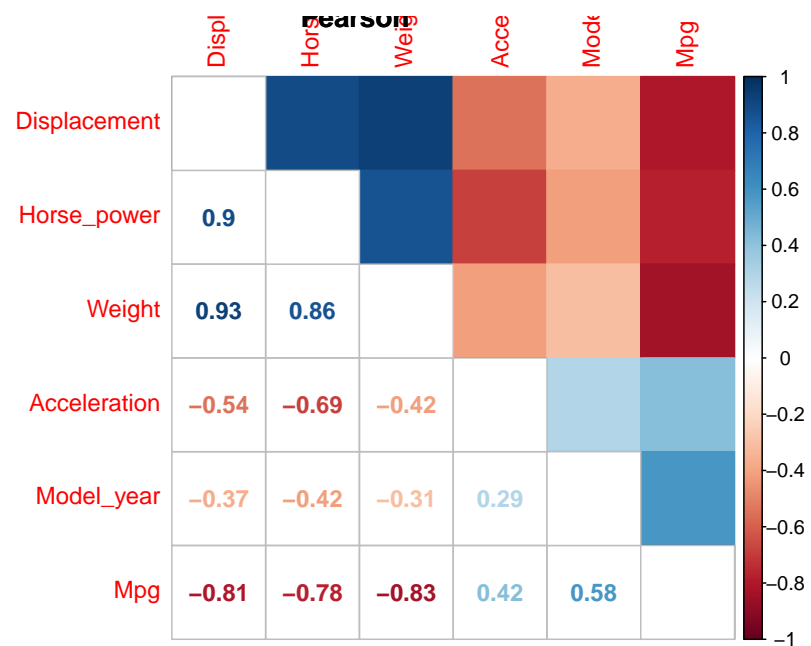


Figura 24

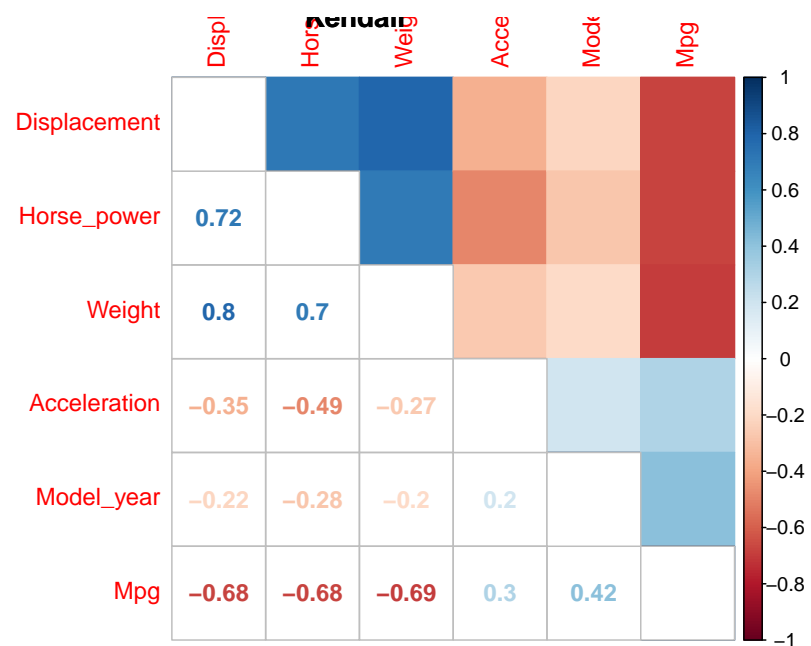


Figura 25

Estas gráficas nos dicen que existe una alta correlación en el dataset, generalmente entre todas las variables (a excepción de Model\_year), pero extremadamente fuerte en las

parejas:

1. Horse\_power & Displacement
2. Weight & Displacement
3. Weight & Horse\_power
4. Acceleration & Horse\_power
5. Mpg & Horse\_power
6. Mpg & Displacement
7. Mpg & Weight

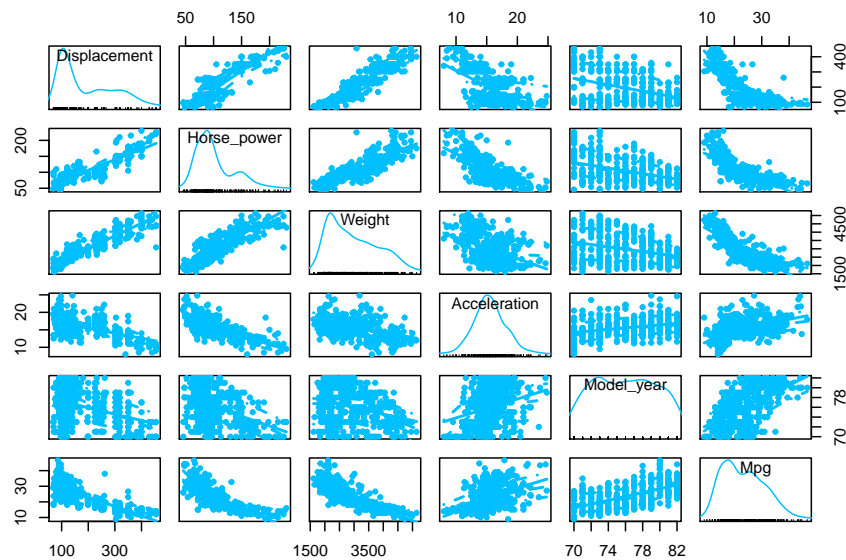


Figura 26

El scatterplot anterior nos muestra mejor la forma de estas correlaciones. Vemos que en todos los casos en los que se da una correlación positiva existe una tendencia lineal entre los datos de ambas variables, y en las negativas una tendencia logarítmica.

Vamos a mostrar algunas parejas con correlación positiva:

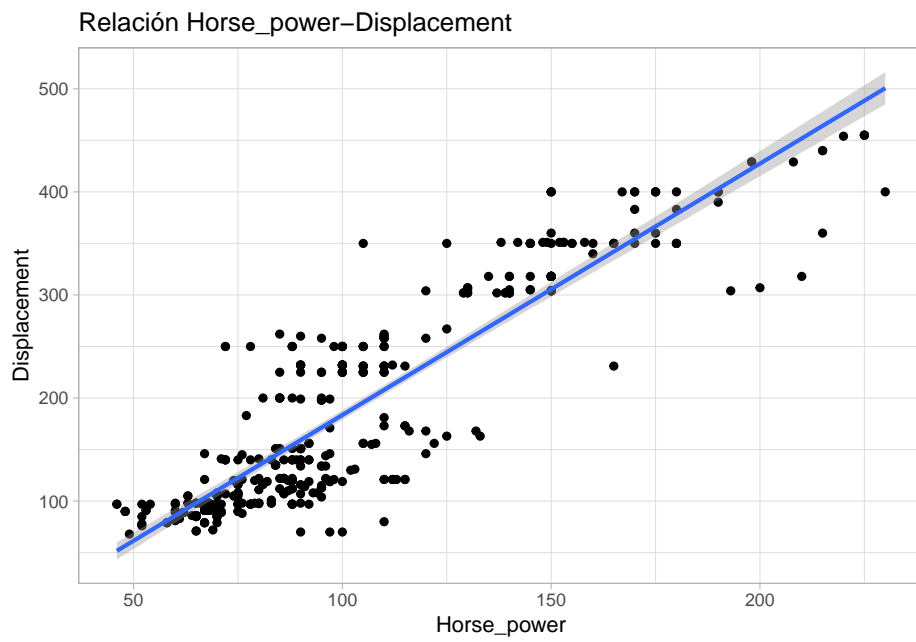


Figura 27

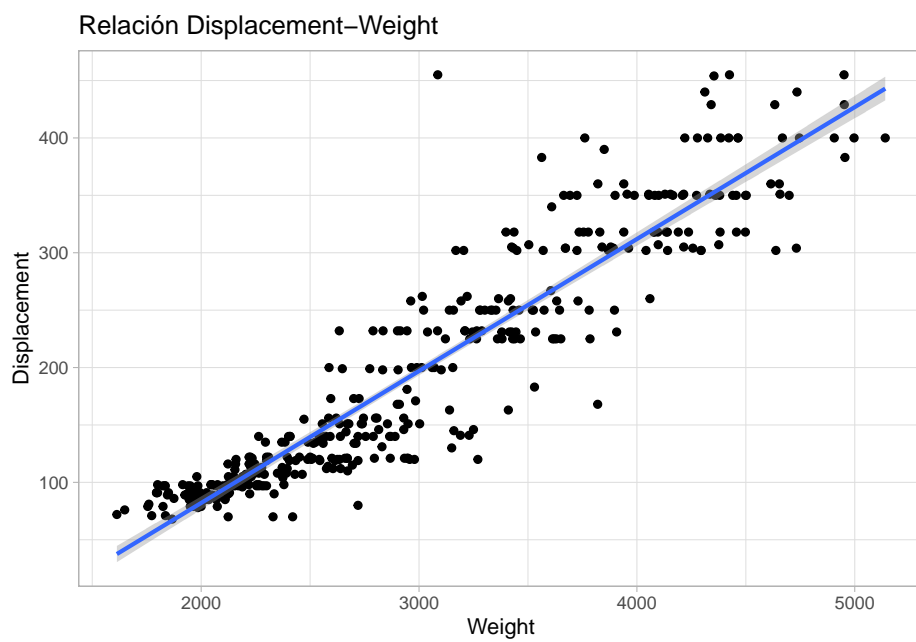


Figura 28

Y otras con correlación negativa:

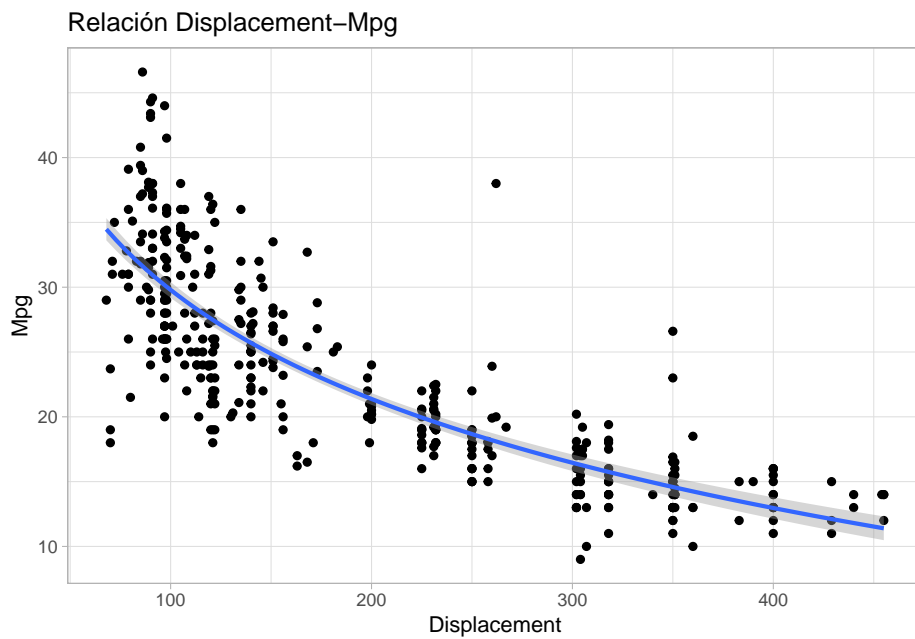


Figura 29

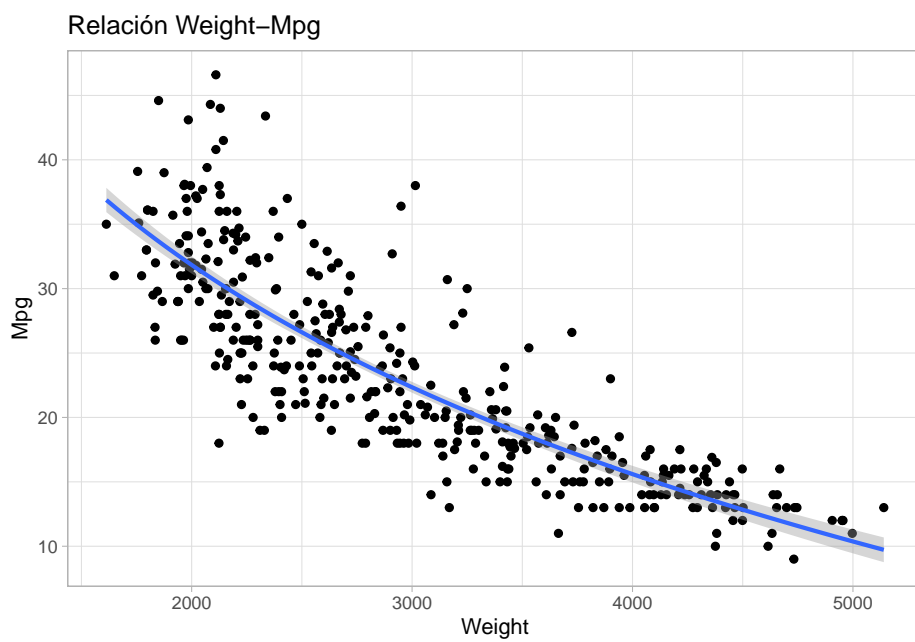


Figura 30

También podemos visualizar los regresores respecto a la salida:

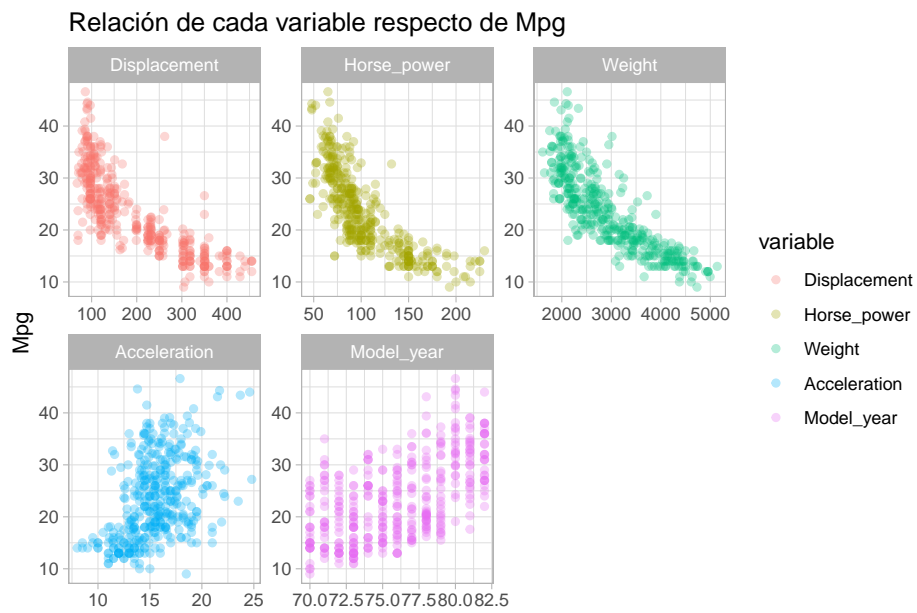


Figura 31

Como habíamos visto, existe alta correlación entre Displacement, Horse\_power, Weight respecto de Mpg.

Haciendo referencia a la hipótesis H.9, Horse\_power podría depender de Displacement y Weight. Parece bastante probable que la potencia de un motor dependa de la cilindrada y el peso que tenga.



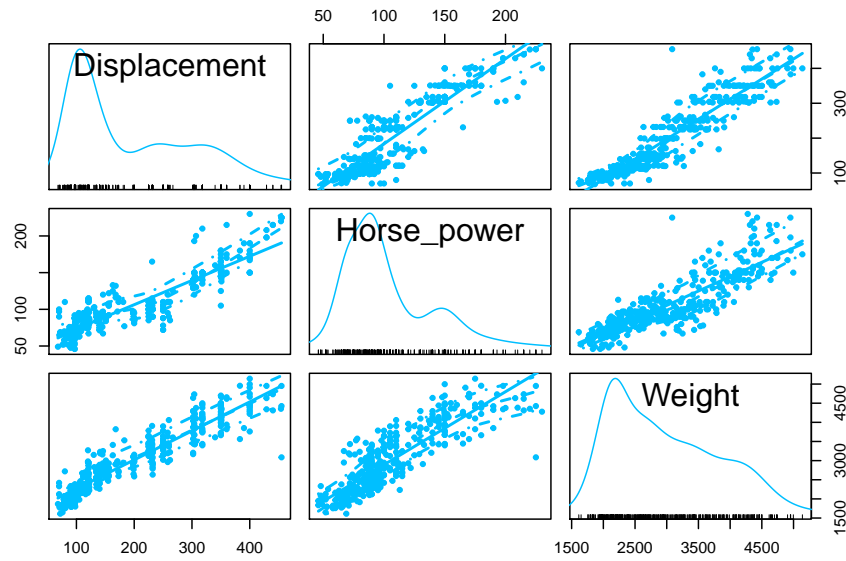


Figura 32

Podemos apreciar como la función de densidad de Horse\_power parece una (“MEDIANIZACIÓN”) de las otras dos.  
Vamos a intentar comprobarlo

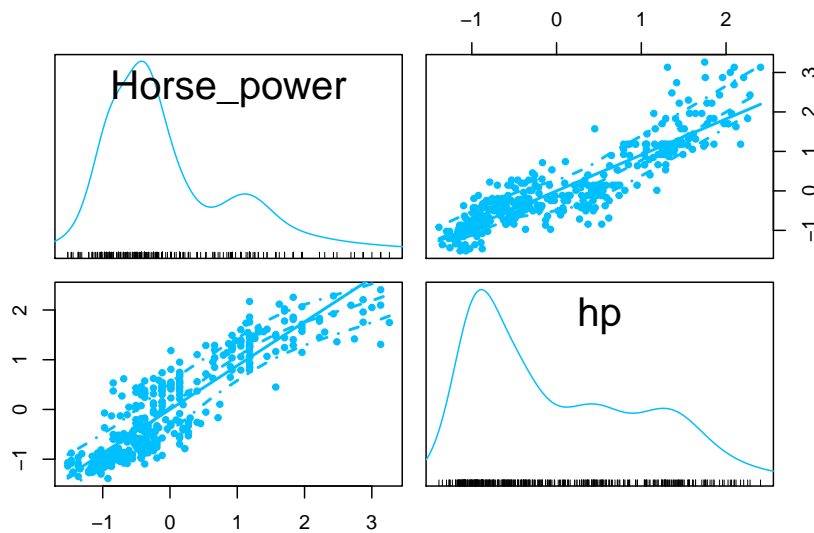


Figura 33

Viendo que no son tan similares como creíamos, buscamos diferentes fórmulas para el cálculo de los caballos de vapor, y vemos que las fórmulas son un poco más complejas y no tenemos exactamente los datos necesarios para utilizarlas (no se descarta que no se puedan deducir, pero no sería un cálculo evidente)

(poner fórmulas [https://www.ajdesigner.com/phphorsepower/horsepower\\_equation\\_trap\\_speed\\_method\\_increase\\_horsepower.php#:~:text=Solving %20for %20the %20change %20in,the %20vehicle](https://www.ajdesigner.com/phphorsepower/horsepower_equation_trap_speed_method_increase_horsepower.php#:~:text=Solving%20for%20the%20change%20in,the%20vehicle))

### 1.2.6. Tratamiento de variables

Para este dataset, al ser casi todas las variables numéricas continuas, existen pocos tratamientos que aplicar.

No tenemos variables categóricas que transformar.

Para añadir interpretabilidad, podríamos agrupar la variable Weight en intervalos, pero puesto que vamos a aplicar regresión sería más conveniente realizarlo con los resultados finales.

### 1.2.7. Ordenaciones

Volvemos a mostrar la cabecera de los datos:

Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
91	70	1955	20.5	71	26.0
232	100	2789	15.0	73	18.0
350	145	4055	12.0	76	13.0
318	140	4080	13.7	78	17.5
113	95	2372	15.0	70	24.0
97	60	1834	19.0	71	27.0

En este caso no es necesario aplicar ninguna reorganización. Cada variable ocupa su propia columna y contiene un único tipo de información, con unidades de observación diferentes. No existe ninguna relación entre variables sobre la información que codifican (en el sentido de que podrían agruparse).

### 1.2.8. Resolución de hipótesis

Nos habíamos planteado las siguientes hipótesis

- **H.1:** Horse\_power puede influir en Mpg: A más potencia, más consumo.

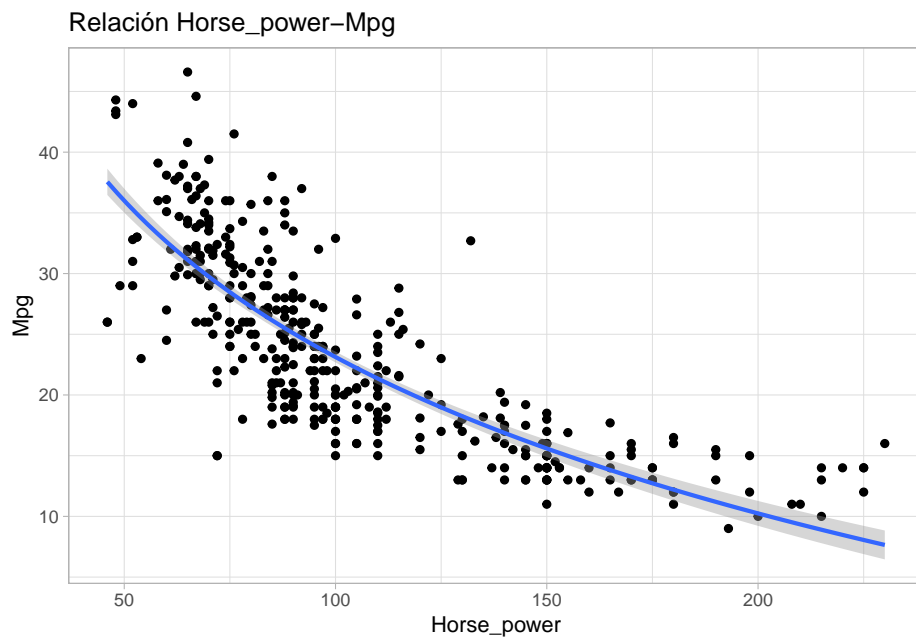


Figura 34

Con el plot y los resultados de la matriz de correlación queda claro que existe una correlación negativa entre estas dos variables. Por tanto, podemos considerar Horse\_power como un buen candidato para la regresión

- **H.2:** Weight debe influir en Mpg: Un coche más pesado debería consumir más idem. a la hipótesis anterior, lo hemos visto anteriormente en la figura X
- **H.3:** Debería haber correlación entre displacement (cilindrada) con horse y acceleration La hemos referenciado anteriormente
- **H.4:** Horse y acceleration podrían estar relacionadas

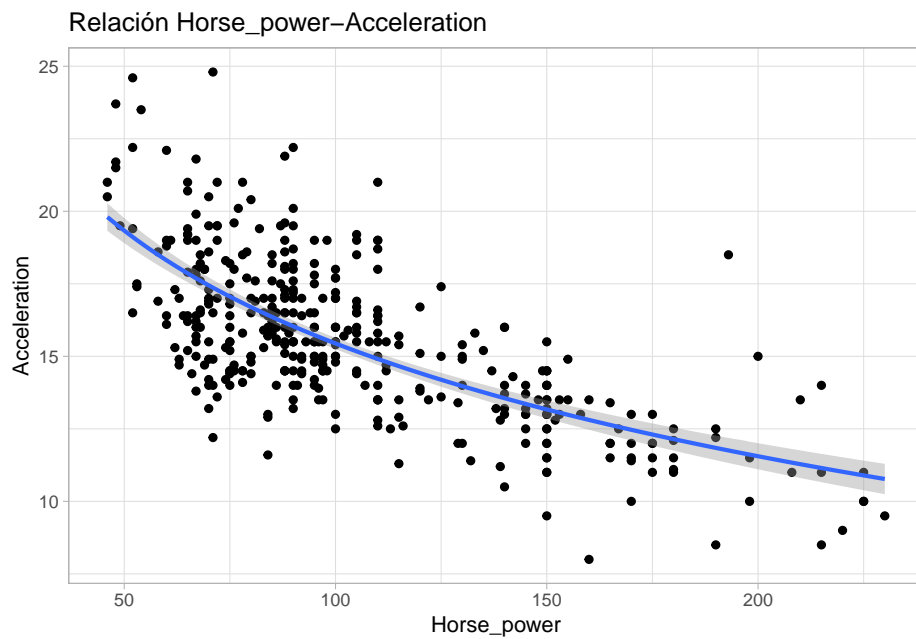


Figura 35

idem. se aprecia una correlación logarítmica entre las dos variables. Similarmente a lo ocurrido con la hipótesis anterior, esto puede ser un problema para nuestro problema de regresión.

- **H.5:** Viendo que contamos con un rango pequeño de años, no debería haber un cambio significativo de prestaciones entre años.

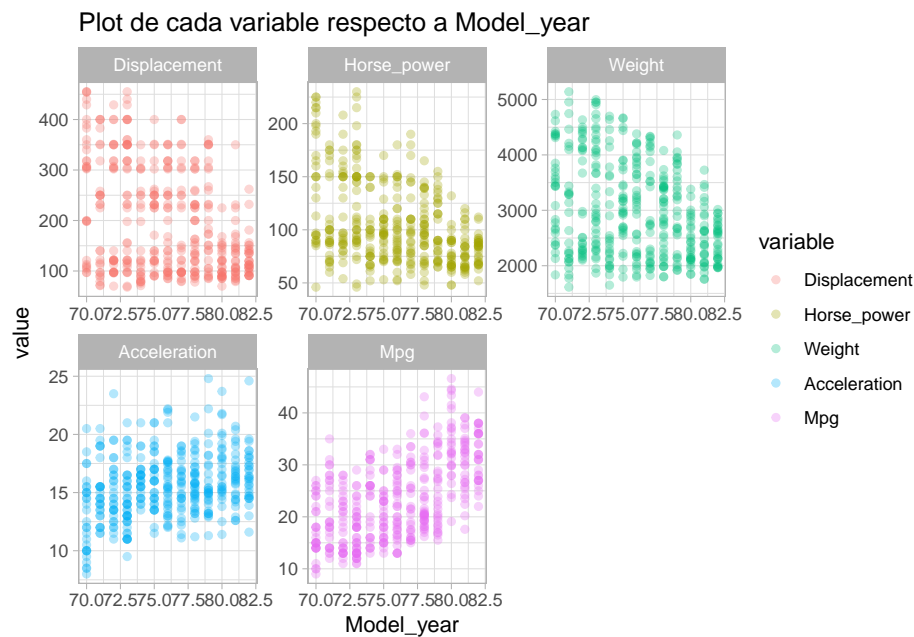


Figura 36

Existe una alta dispersión de los datos en cada una de las variables, pero aún así se aprecia tendencias en las variables. Acceleartion y Mpg tienden a aumentar, y Displacement, Horse\_power y Weight tienden a disminuir. También vemos que la dispersión en las prestaciones de los coches disminuyen ligeramente.

Podemos creer en principio que puede deberse a un decremento del número de instancias con el paso de los años, pero recordamos que en general los datos están repartidos equitativamente

Años: 70 71 72 73 74 75 76 77 78 79 80 81 82  
Conteo: 29 27 28 40 26 30 34 28 36 29 27 28 30

Podemos ver cómo varían los rangos para cada año

Year:	70	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1		97	46	1835	8.0	70	9
2		455	225	4732	20.5	70	27
Year:	71	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1		71	60	1613	11.5	71	12
2		400	180	5140	20.5	71	35
Year:	72	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1		70	54	2100	11.0	72	11
2		429	208	4633	23.5	72	28
Year:	73	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1		68	46	1867	9.5	73	11
2		455	230	4997	21.0	73	29
Year:	74	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1		71	52	1649	13.5	74	13

2	350	150	4699	21.0	74	32
Year: 75	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1	90	53	1795	11.5	75	13
2	400	170	4668	21.0	75	33
Year: 76	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1	85	52	1795	12.0	76	13
2	351	180	4380	22.2	76	33
Year: 77	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1	79	58	1825	11.1	77	15
2	400	190	4335	19.0	77	36
Year: 78	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1	78	48	1800	11.2	78	16.2
2	318	165	4080	21.5	78	43.1
Year: 79	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1	85	65	1915	11.3	79	15.5
2	360	155	4360	24.8	79	37.3
Year: 80	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1	70	48	1845	11.4	80	19.1
2	225	132	3381	23.7	80	46.6
Year: 81	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1	79	58	1755	12.6	81	17.6
2	350	120	3725	20.7	81	39.1
Year: 82	Displacement	Horse_power	Weight	Acceleration	Model_year	Mpg
1	91	52	1965	11.6	82	22
2	262	112	3015	24.6	82	44

- **H.6:** Pero debería existir una tendencia de mejora de prestaciones con los años, incluyendo aumento de Displacement, Horse\_power y Acceleration.

Ciertamente. Se ha comprobado en la hipótesis anterior.

- **H.7:** Model\_year podría no mostrar relación con Mpg: Pese al paso de los años si contamos con diferentes tipos de vehículos (todoterrenos, familiares, deportivos...) podría haber un consumo dispar. (Si existiera tendencia, viendo que los años son de las últimas décadas del siglo XX, podría ir el consumo hacia abajo)

Hemos visto que existe tendencia, lineal con gran dispersión, y positiva.

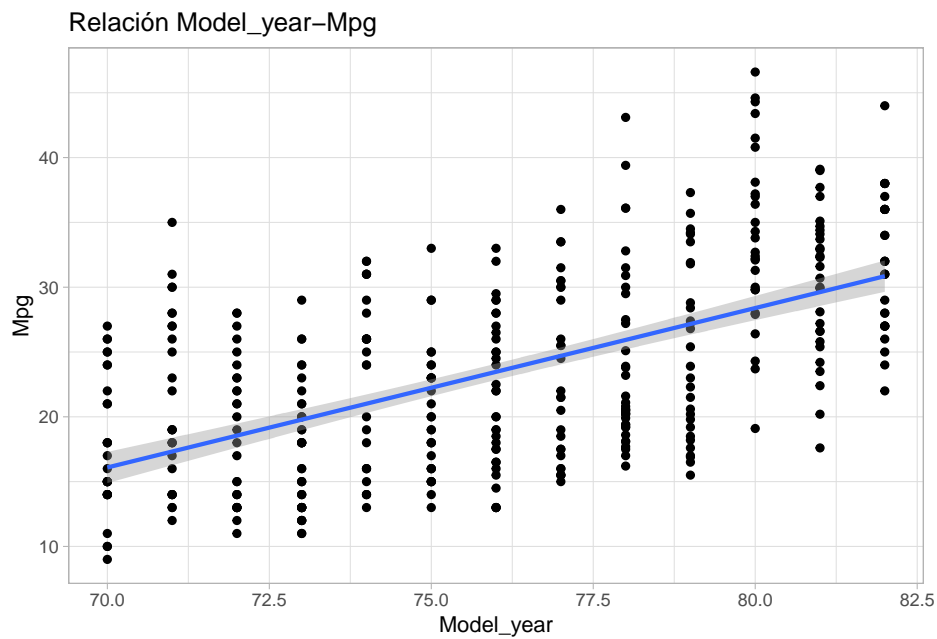


Figura 37

Por desgracia no contamos información sobre los modelos de los coches  
Podemos ver como se ubican los diferentes años en un plot Horse\_power vs Mpg

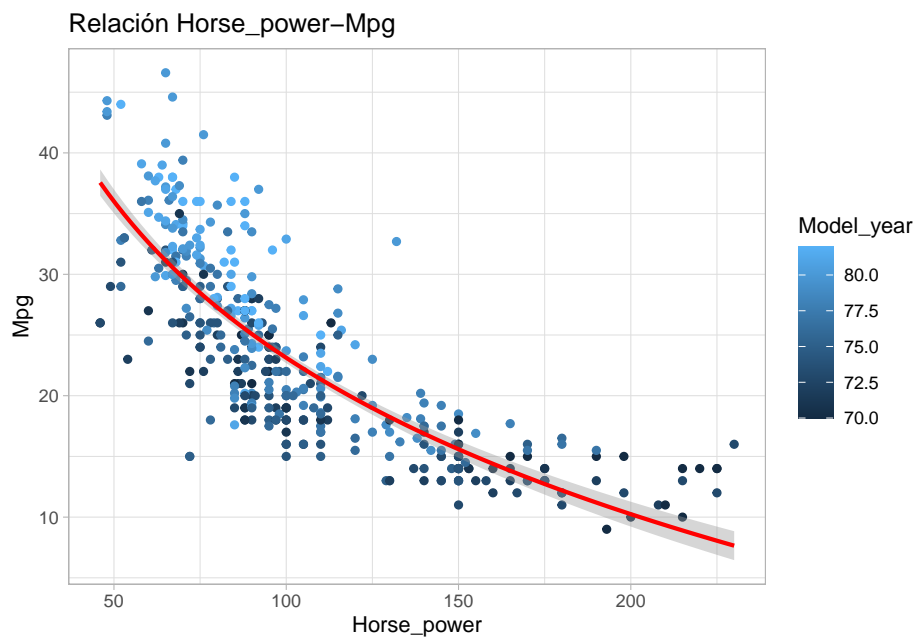


Figura 38

Y vemos que no se puede afirmar la hipótesis, los coches están entremezclados por diferentes años

- **H.8:** Esta última hipótesis se puede aplicar al resto de variables, indicándonos que `Model_year` no debería tener relevancia para este problema de regresión.

No podemos afirmar la hipótesis anterior y por consiguiente esta tampoco.

- **H.9:** `Horse_power` podría depender de las variables `Displacement` y `Weight`

Lo hemos comentado anteriormente.

### 1.3. Conclusiones

Como conclusiones podemos decir que tenemos un dataset altamente correlacionado, distribuido de forma no normal pero con la información bien representada. Existen relaciones fuertes entre las variables de entrada y de las de salida para la regresión que probablemente nos ayuden a solucionar con facilidad el problema.

Aunque no hemos descubierto los tipos de distribución que siguen nuestras variables, por si quisiéramos transformarlas a una normal, podemos sin ninguna duda aplicar una estandarización de los datos (puesto que sabemos que no afecta negativamente al problema de regresión) siempre y cuando lo tengamos en cuenta a la hora de analizar los resultados.

Se nos pide elegir 5 regresores para la regresión y contamos exactamente con ese número, por lo que no podemos descartar ninguna variable. Aún así, hemos visto que tenemos algunas variables más interesantes que otras. Variables correladas con la salida nos aumentan las posibilidades de obtener un buen regresor, pero debemos evitar usar variables correladas entre sí para evitar la multicolinealidad. Sería conveniente evitarla para aumentar la interpretabilidad del modelo, pero la potencia en sí de este no cambia. (<https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/#:~:text=Multicollinearity> (referenciar esta frase en el apartado de regresión))



## 2. Técnicas de Regresión

Recordamos que la descripción de los datos se encuentra en el apartado 1.1.

Como se comentó en el apartado de EDA:

“Se nos pide elegir 5 regresores para la regresión y contamos exactamente con ese número, por lo que no podemos descartar ninguna variable. Aún así, hemos visto que tenemos algunas variables más interesantes que otras. Variables correladas con la salida nos aumentan las posibilidades de obtener un buen regresor, pero debemos evitar usar variables correladas entre sí para evitar la multicolinealidad. Sería conveniente evitarla para aumentar la interpretabilidad del modelo, pero la potencia en sí de este no cambia.”

Graficamos la relación de cada variable respecto a la salida

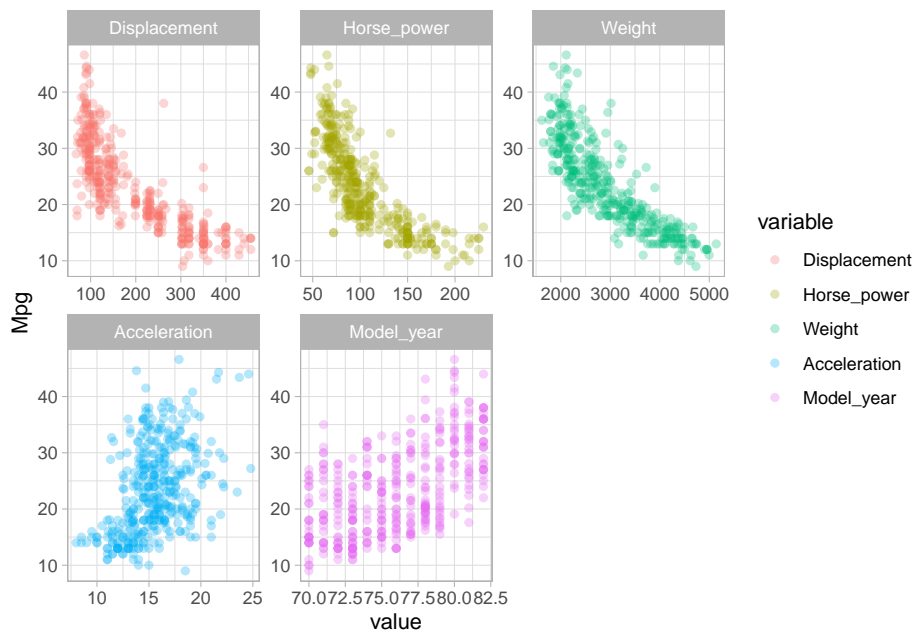


Figura 39

Como dijimos, se aprecia alta correlación entre Displacement, Horse\_power, Weight respecto de la salida, probablemente de forma logarítmica.

Las matrices nos correlación nos confirman esta idea (con coeficientes de Pearson y Kendall)

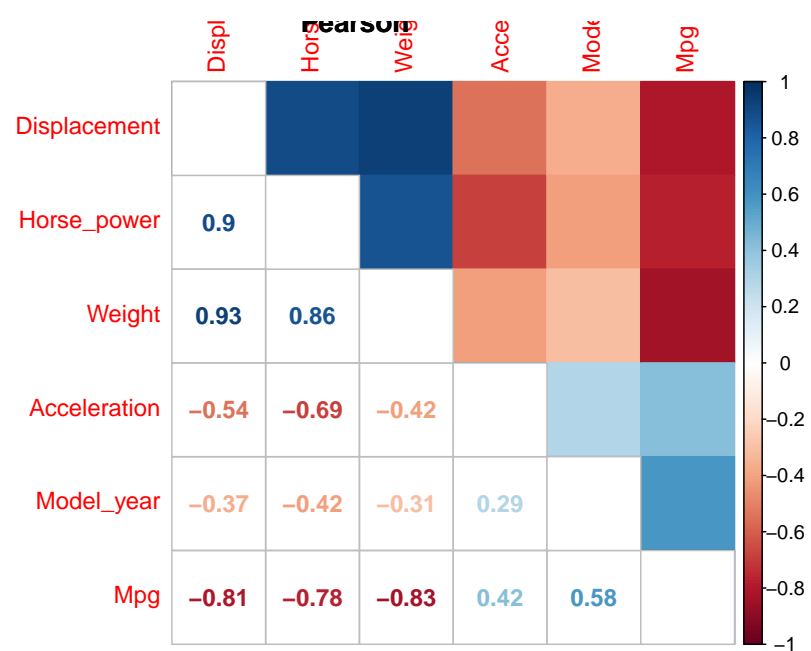


Figura 40

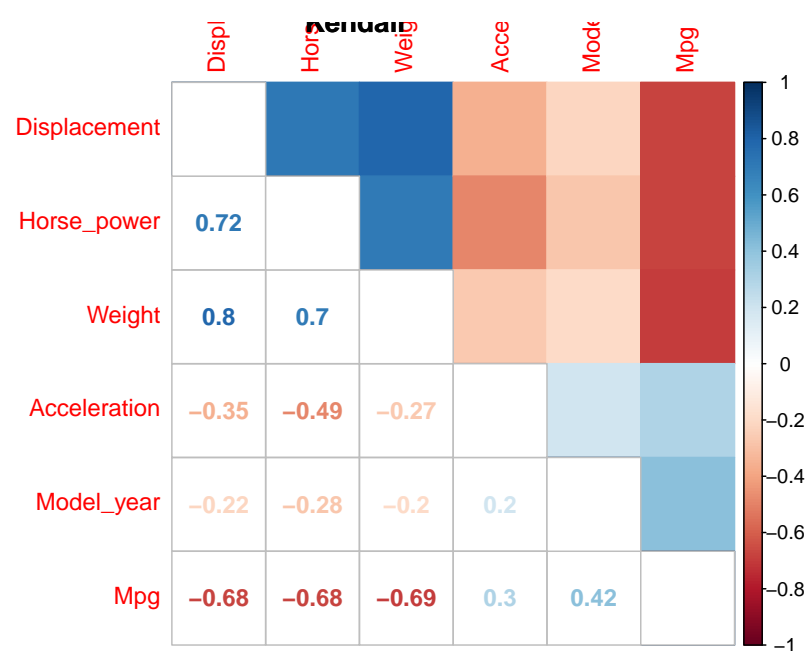


Figura 41

Por tanto, si las ordenáramos por cuáles parecen ser más prometedoras, tendríamos: Weight > Displacement > Horse\_power > Model\_year > Acceleration

También tenemos que tener en cuenta que las tres primeras variables están correladas entre sí.

## 2.1. Ajustes de regresión lineal univariabes

Vamos a analizar un ajuste con cada una de las características:

```
Call:
lm(formula = Mpg ~ Weight, data = auto)

Residuals:
    Min       1Q   Median       3Q      Max
-11.9736  -2.7556  -0.3358   2.1379  16.5194

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.216524    0.798673   57.87  <2e-16 ***
Weight       -0.007647    0.000258  -29.64  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.333 on 390 degrees of freedom
Multiple R-squared:  0.6926,    Adjusted R-squared:  0.6918
F-statistic: 878.8 on 1 and 390 DF,  p-value: < 2.2e-16
```

"-----"

```
Call:
lm(formula = Mpg ~ Displacement, data = auto)

Residuals:
    Min       1Q   Median       3Q      Max
-12.9170  -3.0243  -0.5021   2.3512  18.6128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.12064    0.49443   71.03  <2e-16 ***
Displacement -0.06005    0.00224  -26.81  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.635 on 390 degrees of freedom
Multiple R-squared:  0.6482,    Adjusted R-squared:  0.6473
F-statistic: 718.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

"-----"

```
Call:
lm(formula = Mpg ~ Horse_power, data = auto)
```

Residuals:

Min	1Q	Median	3Q	Max
-13.5710	-3.2592	-0.3435	2.7630	16.9240

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	39.935861	0.717499	55.66	<2e-16 ***
Horse_power	-0.157845	0.006446	-24.49	<2e-16 ***

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

Residual standard error: 4.906 on 390 degrees of freedom  
Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049  
F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

"-----"

Call:

lm(formula = Mpg ~ Model\_year, data = auto)

Residuals:

Min	1Q	Median	3Q	Max
-12.0212	-5.4411	-0.4412	4.9739	18.2088

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-70.01167	6.64516	-10.54	<2e-16 ***
Model_year	1.23004	0.08736	14.08	<2e-16 ***

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

Residual standard error: 6.363 on 390 degrees of freedom  
Multiple R-squared: 0.337, Adjusted R-squared: 0.3353  
F-statistic: 198.3 on 1 and 390 DF, p-value: < 2.2e-16

"-----"

Call:

lm(formula = Mpg ~ Acceleration, data = auto)

Residuals:

Min	1Q	Median	3Q	Max
-17.989	-5.616	-1.199	4.801	23.239

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.8332	2.0485	2.359	0.0188 *
Acceleration	1.1976	0.1298	9.228	<2e-16 ***

```

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

Residual standard error: 7.08 on 390 degrees of freedom
Multiple R-squared:  0.1792,    Adjusted R-squared:  0.1771 
F-statistic: 85.15 on 1 and 390 DF,  p-value: < 2.2e-16

```

Al ser univariable, no es necesario fijarse en el estadístico F por ahora. Para ver el potencial de la variable, debemos darle importancia al p-valor (comprobar de que sea lo suficientemente bajo), y posteriormente ver el R2 para averiguar el porcentaje de la salida explicada.

En base a los resultados vemos que el test de correlación nos había ayudado correctamente: de forma individual todas las variables tienen dependencia lineal, y el orden de calidad coincide con el orden de fuerza en las correlaciones.

Guardamos el modelo aditivo hasta ahora

Ya con el uso de la variable Weight vemos que podemos explicar un ~69% de la salida, un buen valor de partida. Graficamos el ajuste:

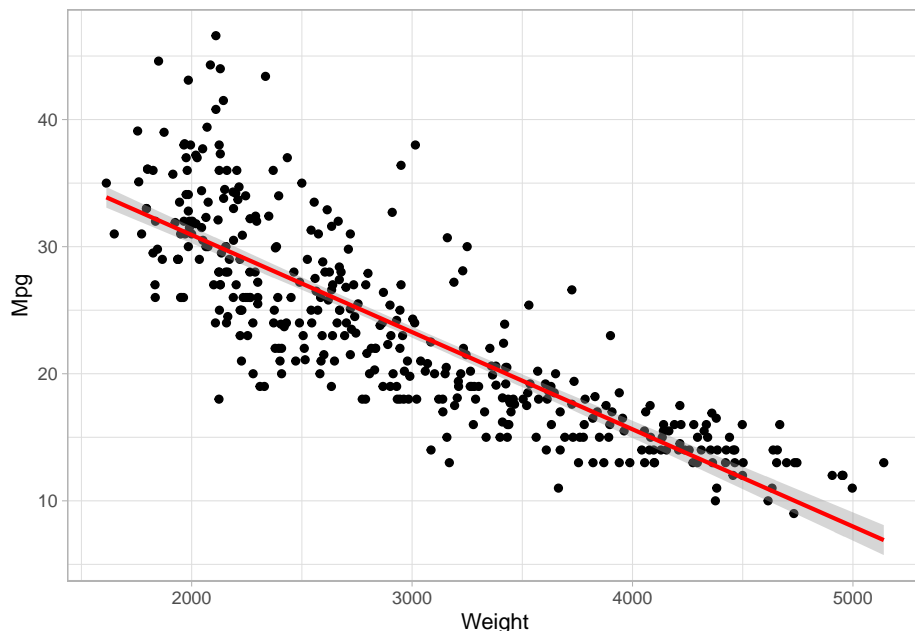


Figura 42

Y vemos sus coeficientes:

	2.5 %	97.5 %
(Intercept)	44.646282308	47.78676679
Weight	-0.008154515	-0.00714017

Aunque los valores del intervalo del coeficiente de Weight sea bajo, vemos que no incluye el cero (y con el p-valor obtenido anteriormente, lo podemos asegurar con bastante

certeza). Probablemente la razón de estos coeficientes tan pequeños es que los datos no están estandarizados (se podría hacer perfectamente, se han dejado con sus rangos normales para interpretarlos mejor) y los valores de las unidades de medida son bastante diferentes (hablamos de rangos de [9.0,46.6] en Mpg frente a [1613,5140] en Weight)

Ya con esto podemos intentar interpretar un poco los datos, tendríamos por ahora la fórmula de regresión lineal: (poner fórmula)

(INTERPRETAR ?) La fórmula nos indica que por cada unidad de peso el Mpg decrementa ?

## 2.2. Ajustes de regresión lineal multivariable

Aplicamos un método descendente

Call:

```
lm(formula = Mpg ~ ., data = auto)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.5211	-2.3920	-0.1036	2.0312	14.2874

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-1.544e+01	4.677e+00	-3.300	0.00106	**
Displacement	2.782e-03	5.462e-03	0.509	0.61082	
Horse_power	1.020e-03	1.376e-02	0.074	0.94095	
Weight	-6.874e-03	6.653e-04	-10.333	< 2e-16	***
Acceleration	9.032e-02	1.019e-01	0.886	0.37599	
Model_year	7.541e-01	5.261e-02	14.334	< 2e-16	***

---

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

Residual standard error: 3.435 on 386 degrees of freedom

Multiple R-squared: 0.8088, Adjusted R-squared: 0.8063

F-statistic: 326.5 on 5 and 386 DF, p-value: < 2.2e-16

El p-valor del F estadístico nos dice que al menos hay una variable (realmente ya lo sabíamos de los ajustes univariantes) con dependencia lineal.

Vemos que hay 3 variables con mal p-valor, empezamos quitando la que lo tiene más alto, Horse\_power

Call:

```
lm(formula = Mpg ~ . - Horse_power, data = auto)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.5182	-2.3948	-0.1085	2.0405	14.2908

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.527e+01  4.106e+00  -3.719 0.000229 ***
Displacement  2.874e-03  5.310e-03   0.541 0.588651
Weight       -6.852e-03  5.967e-04 -11.483 < 2e-16 ***
Acceleration  8.555e-02  7.885e-02   1.085 0.278595
Model_year    7.532e-01  5.118e-02  14.717 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 3.431 on 387 degrees of freedom
Multiple R-squared:  0.8088,    Adjusted R-squared:  0.8068
F-statistic: 409.2 on 4 and 387 DF,  p-value: < 2.2e-16

```

El F estadístico está correcto, y seguimos teniendo variables con p-valor grande, quitamos Displacement

```

Call:
lm(formula = Mpg ~ . - Horse_power - Displacement, data = auto)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-8.6749 -2.3528 -0.1082  2.0168 14.3022

```

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.936555   4.055512  -3.683 0.000263 ***
Weight       -0.006554   0.000230 -28.502 < 2e-16 ***
Acceleration  0.066359   0.070361   0.943 0.346204
Model_year    0.748446   0.050366  14.860 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 3.428 on 388 degrees of freedom
Multiple R-squared:  0.8086,    Adjusted R-squared:  0.8071
F-statistic: 546.5 on 3 and 388 DF,  p-value: < 2.2e-16

```

idem. a lo anterior, quitamos Acceleration.

```

Call:
lm(formula = Mpg ~ . - Horse_power - Displacement - Acceleration,
    data = auto)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-8.8505 -2.3014 -0.1167  2.0367 14.3555

```

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)

```

```

(Intercept) -1.435e+01  4.007e+00  -3.581 0.000386 ***
Weight      -6.632e-03  2.146e-04 -30.911 < 2e-16 ***
Model_year   7.573e-01  4.947e-02  15.308 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.427 on 389 degrees of freedom
Multiple R-squared:  0.8082,    Adjusted R-squared:  0.8072
F-statistic: 819.5 on 2 and 389 DF,  p-value: < 2.2e-16

```

El estadístico F sigue bien, y los p-valores de las variables son extremadamente bajos. Nos fijamos en el R2 y vemos que ha subido considerablemente (un 10%) respecto al univariable, por lo que este sería nuestro modelo aditivo por ahora.

A partir de ahora hay que tener cuidado si el R2 sigue aumentando, hay que evitar el overfitting en el modelo.

### 2.3. Inserción de interacciones

Del modelo aditivo solo nos han quedado dos regresores, así que probamos a incluirlos como interacción.

```

Call:
lm(formula = Mpg ~ +Weight * Model_year, data = auto)

Residuals:
    Min       1Q   Median       3Q      Max
-8.0397 -1.9956 -0.0983  1.6525 12.9896

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.105e+02  1.295e+01  -8.531 3.30e-16 ***
Weight          2.755e-02  4.413e-03   6.242 1.14e-09 ***
Model_year      2.040e+00  1.718e-01  11.876 < 2e-16 ***
Weight:Model_year -4.579e-04  5.907e-05  -7.752 8.02e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.193 on 388 degrees of freedom
Multiple R-squared:  0.8339,    Adjusted R-squared:  0.8326
F-statistic: 649.3 on 3 and 388 DF,  p-value: < 2.2e-16

```

El F estadístico sigue bien y los p-valores son bajos, el nuevo R2 ha mejorado un 3 %, así que no es demasiado para considerar un overfitting. Probablemente más de un 90 % sería preocupante, pero también tenemos que tener en cuenta que las variables están fuertemente correladas con la salida.

Podríamos probar a añadir alguna interacción más con alguna variable que no hubiera entrado en el modelo aditivo, pero no se espera que mejore:



```

Call:
lm(formula = Mpg ~ +Weight * Model_year + Acceleration * Displacement,
    data = auto)

Residuals:
    Min       1Q   Median       3Q      Max
-7.3130 -1.8670 -0.0426  1.6109 12.2499

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.131e+02  1.321e+01  -8.564 2.65e-16 ***
Weight          2.456e-02  4.693e-03   5.234 2.73e-07 ***
Model_year      1.907e+00  1.769e-01  10.778 < 2e-16 ***
Acceleration     7.273e-01  1.282e-01   5.671 2.79e-08 ***
Displacement    3.605e-02  8.673e-03   4.157 3.98e-05 ***
Weight:Model_year -4.054e-04  6.281e-05  -6.454 3.29e-10 ***
Acceleration:Displacement -2.953e-03  6.219e-04  -4.748 2.91e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.075 on 385 degrees of freedom
Multiple R-squared:  0.8472,    Adjusted R-squared:  0.8448
F-statistic: 355.7 on 6 and 385 DF,  p-value: < 2.2e-16

```

A pesar de nuestra suposición los p-valores son válidos y el R2 aumenta un 1%. Es cuestionable si el aumento de la complejidad del modelo merece con este incremento de R2. Por simplificar vamos a quedarnos con el modelo aditivo anterior y probar con otra interacción.

Podemos probar combinando la variable Acceleration con una de que teníamos (Weight y Model\_year)

```

Call:
lm(formula = Mpg ~ +Weight * Model_year + Acceleration * Weight,
    data = auto)

Residuals:
    Min       1Q   Median       3Q      Max
-7.4473 -1.7994 -0.0496  1.4790 12.1258

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.230e+02  1.298e+01  -9.480 < 2e-16 ***
Weight          2.971e-02  4.419e-03   6.722 6.47e-11 ***
Model_year      1.926e+00  1.742e-01  11.055 < 2e-16 ***
Acceleration     1.341e+00  2.323e-01   5.772 1.61e-08 ***
Weight:Model_year -4.078e-04  6.197e-05  -6.581 1.53e-10 ***
Weight:Acceleration -3.808e-04  7.537e-05  -5.052 6.76e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 3.061 on 386 degrees of freedom  
 Multiple R-squared: 0.8482, Adjusted R-squared: 0.8462  
 F-statistic: 431.4 on 5 and 386 DF, p-value: < 2.2e-16

Call:

```
lm(formula = Mpg ~ +Weight * Model_year + Acceleration * Model_year,
    data = auto)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.8674	-1.9539	-0.0617	1.7397	12.3964

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-4.046e+01	2.833e+01	-1.428	0.15401
Weight	2.523e-02	4.881e-03	5.170	3.76e-07 ***
Model_year	1.072e+00	3.704e-01	2.895	0.00401 **
Acceleration	-3.956e+00	1.268e+00	-3.120	0.00195 **
Weight:Model_year	-4.263e-04	6.475e-05	-6.584	1.49e-10 ***
Model_year:Acceleration	5.476e-02	1.663e-02	3.293	0.00108 **

---

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

Residual standard error: 3.117 on 386 degrees of freedom  
 Multiple R-squared: 0.8426, Adjusted R-squared: 0.8406  
 F-statistic: 413.2 on 5 and 386 DF, p-value: < 2.2e-16

Y entre los dos nos podríamos quedar con el primero por tener mejores p-valores y un mejor R2. Aun así, el incremento es pequeño respecto a nuestro modelo aditivo.

La fórmula del modelo aditivo que llevamos por ahora es:

Y la graficamos:

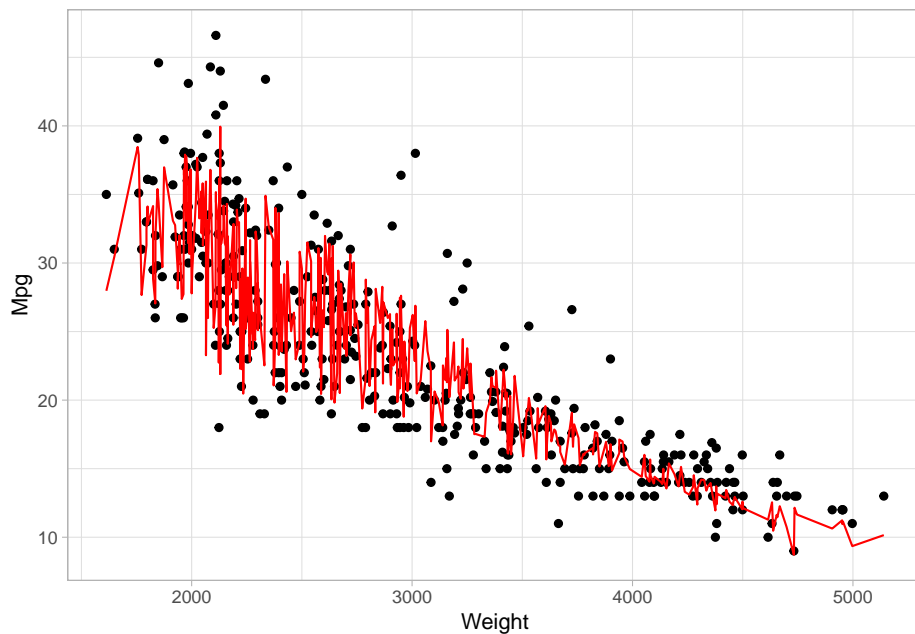


Figura 43

Por el uso multivariable la línea de regresión que se forma podría indicarnos un posible overfitting en el modelo, vamos a dejarlo por ahora e intentar solucionarlo con el modelo no lineal.

## 2.4. Ajustes de regresión no lineal

Habíamos dicho que las gráficas nos mostraban una tendencia logarítmica, vamos a incluir la de Weight en nuestro modelo aditivo

Call:

```
lm(formula = Mpg ~ +Weight * Model_year + Acceleration * Weight +  
  I(log(Weight)), data = auto)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.6734	-1.7933	-0.0576	1.3154	12.1716

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	1.191e+02	4.120e+01	2.891	0.00406	**
Weight	2.842e-02	4.227e-03	6.723	6.44e-11	***
Model_year	1.638e+00	1.728e-01	9.480	< 2e-16	***
Acceleration	7.236e-01	2.435e-01	2.972	0.00315	**
I(log(Weight))	-3.028e+01	4.914e+00	-6.162	1.81e-09	***
Weight:Model_year	-2.971e-04	6.186e-05	-4.803	2.24e-06	***

```

Weight:Acceleration -1.775e-04  7.919e-05  -2.241  0.02559 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.924 on 385 degrees of freedom
Multiple R-squared:  0.8618,    Adjusted R-squared:  0.8597
F-statistic: 400.2 on 6 and 385 DF,  p-value: < 2.2e-16

```

El estadístico F está bien y los p-valores también, aunque el de la interacción Weight-Acceleration es alto comparado con el resto (aún así sigue siendo aceptable).

Como el R2 ha subido, por ver si mejora, vamos a quitar esta interacción.

```

Call:
lm(formula = Mpg ~ +Weight * Model_year + I(log(Weight)), data = auto)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-8.7501 -1.7470 -0.0725  1.3122 12.6776

```

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.715e+02  3.810e+01   4.501 8.98e-06 ***
Weight         2.522e-02  4.119e-03   6.123 2.25e-09 ***
Model_year     1.572e+00  1.708e-01   9.202 < 2e-16 ***
I(log(Weight)) -3.540e+01  4.538e+00 -7.800 5.82e-14 ***
Weight:Model_year -2.701e-04  6.003e-05 -4.499 9.04e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 2.972 on 387 degrees of freedom
Multiple R-squared:  0.8565,    Adjusted R-squared:  0.855
F-statistic: 577.3 on 4 and 387 DF,  p-value: < 2.2e-16

```

Hemos empeorado un 0.5%, bastante poco, y el modelo es más simple. La dejamos quitada.

Podemos mostrarlo en un gráfico:

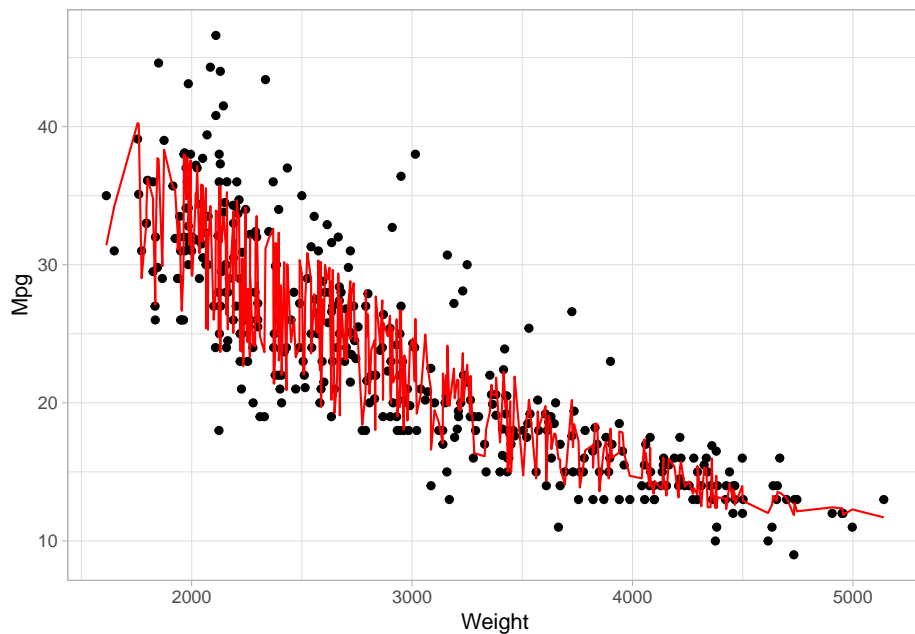


Figura 44

Esta gráfica nos indica que con casi toda probabilidad se está generando sobreajuste, se ve necesario simplificar el modelo.

Si quitamos la otra interacción:

Call:

```
lm(formula = Mpg ~ Weight + Model_year + I(log(Weight)), data = auto)
```

Residuals:

Min	1Q	Median	3Q	Max
-9.3384	-1.7476	-0.2122	1.5322	13.2812

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	284.287315	29.392946	9.672	< 2e-16 ***
Weight	0.007772	0.001420	5.473	7.97e-08 ***
Model_year	0.828693	0.044506	18.620	< 2e-16 ***
I(log(Weight))	-43.590633	4.258803	-10.235	< 2e-16 ***

---

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

Residual standard error: 3.045 on 388 degrees of freedom

Multiple R-squared: 0.849, Adjusted R-squared: 0.8478

F-statistic: 727 on 3 and 388 DF, p-value: < 2.2e-16

No hemos perdido apenas R2, mostramos la gráfica:

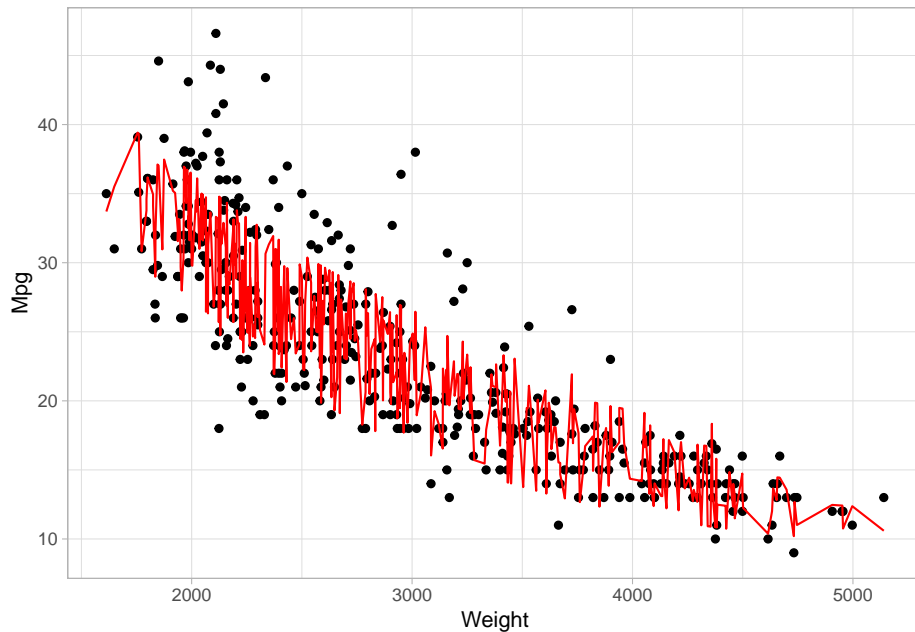


Figura 45

Seguimos con el mismo problema, probablemente se deba a una de las variables. Quitamos `Model_year` por tener poca correlación con la variable de salida:

Call:

```
lm(formula = Mpg ~ Weight + I(log(Weight)), data = auto)
```

Residuals:

Min	1Q	Median	3Q	Max
-12.5329	-2.7031	-0.4016	1.7038	16.0835

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	263.812407	40.366256	6.535	1.99e-10 ***
Weight	0.002582	0.001914	1.349	0.178
I(log(Weight))	-31.166013	5.780558	-5.392	1.21e-07 ***

---

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

Residual standard error: 4.185 on 389 degrees of freedom

Multiple R-squared: 0.714, Adjusted R-squared: 0.7125

F-statistic: 485.6 on 2 and 389 DF, p-value: < 2.2e-16

El p-valor de `Weight` nos indica que hay que quitarla, y al no estar incluida ninguna interacción, no es un término de jerarquía, por lo que podemos hacerlo. Se puede porque la variable sigue siendo independiente, solamente no está modelada de forma lineal, sino logarítmicamente.

```

Call:
lm(formula = Mpg ~ I(log(Weight)), data = auto)

Residuals:
    Min       1Q   Median       3Q      Max
-12.4315  -2.6752  -0.2888   1.9429  16.0136

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   209.9433     6.0002   34.99  <2e-16 ***
I(log(Weight)) -23.4317     0.7534  -31.10  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.189 on 390 degrees of freedom
Multiple R-squared:  0.7127,    Adjusted R-squared:  0.7119
F-statistic: 967.3 on 1 and 390 DF,  p-value: < 2.2e-16

```

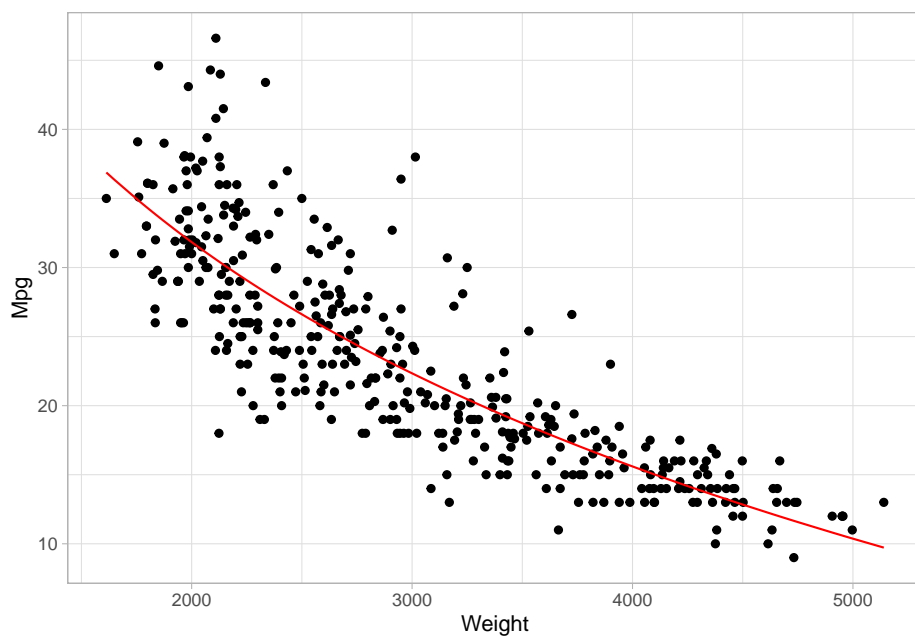


Figura 46

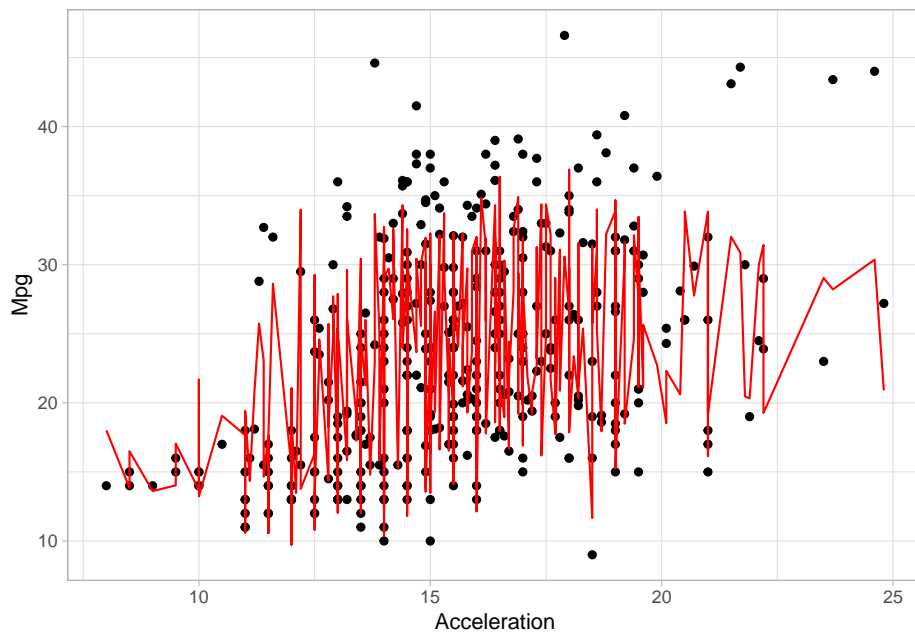


Figura 47

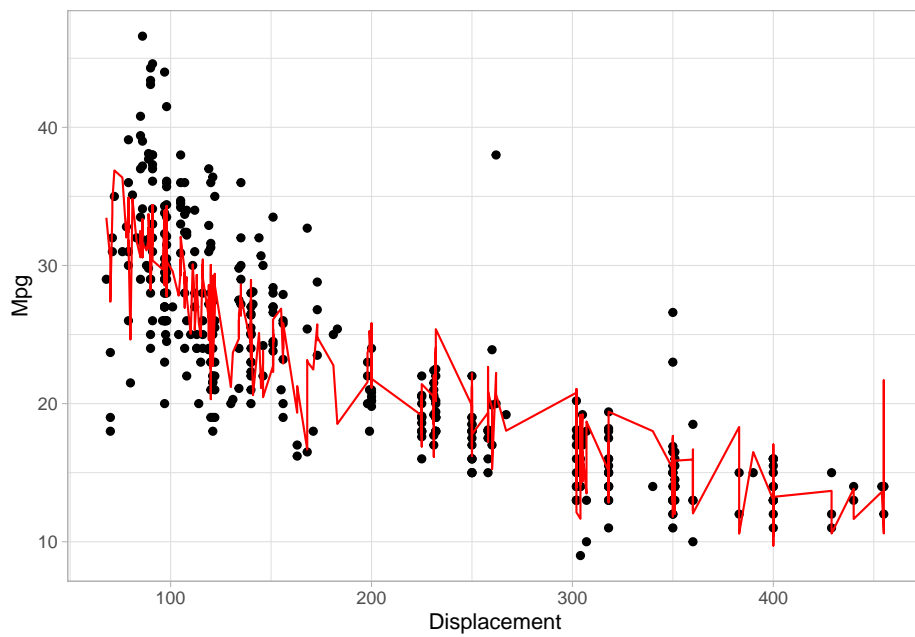


Figura 48

Vemos un empeoramiento significativo en la calidad de  $R^2$  respecto al modelo multivariable, pero la forma del modelo no está tan ajustada a los datos y parece sensato



mantenerlo así.

Aún así, no resulta lógico intentar predecir el Mpg de un coche únicamente en base al peso, alguna de las otras variables deberían ayudarnos en la predeción. Por ejemplo, alguna característica del motor, como la cilindrada o los caballos de vapor.

Para resumir, mostramos el modelo con mejor R2 tras hacer múltiples pruebas, e intentando evitar un overfitting:

Call:

```
lm(formula = Mpg ~ Acceleration + I(log(Weight)) + I(log(Displacement)),
    data = auto)
```

Residuals:

Min	1Q	Median	3Q	Max
-12.9074	-2.6174	-0.4104	1.9500	16.5596

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	171.61778	12.14751	14.128	< 2e-16 ***
Acceleration	0.19717	0.08914	2.212	0.0276 *
I(log(Weight))	-16.94003	2.27727	-7.439	6.59e-13 ***
I(log(Displacement))	-3.19963	1.26881	-2.522	0.0121 *

---

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

Residual standard error: 4.104 on 388 degrees of freedom

Multiple R-squared: 0.7256, Adjusted R-squared: 0.7235

F-statistic: 342.1 on 3 and 388 DF, p-value: < 2.2e-16

Los p-valores no son muy fuertes, pero siguen siendo aceptables, y gráficamente el modelo se ve un poco mejor:

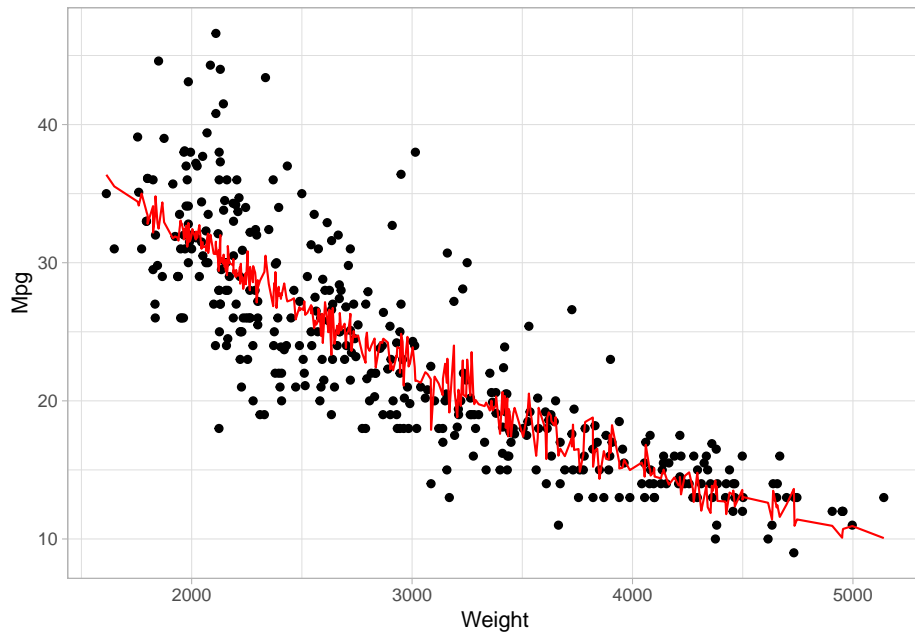


Figura 49

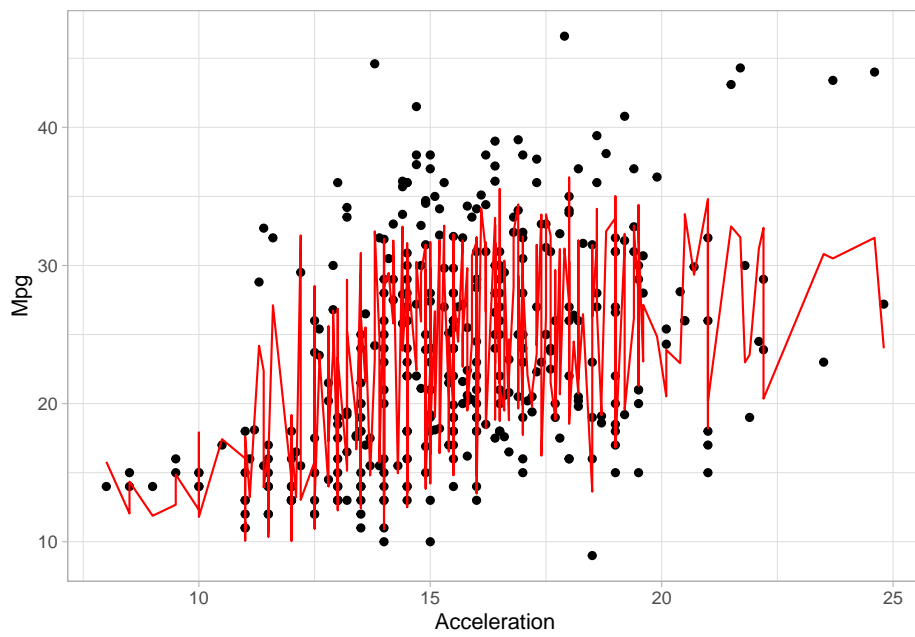


Figura 50

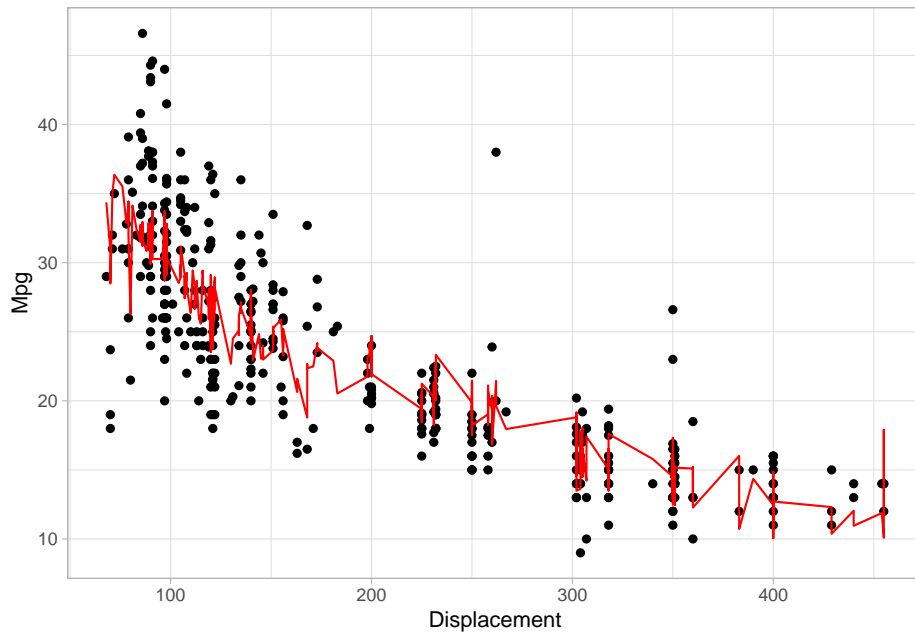


Figura 51

Pensando en el problema, y tras el análisis hecho en el apartado de EDA, creemos que usar `Model_year` para predecir `Mpg` no parece buena idea. La gráfica de la variable nos muestran mucha dispersión en los datos y, aunque sí se ve una cierta tendencia lineal, no parece suficiente para usarla. Claramente nos ajusta mejor los datos pero parece que nos estamos pegando a ellos.

De cara a comprobar este razonamiento en el cross-validation, vamos a guardar dos modelos: - Modelo con mejor  $R^2$   $Mpg \sim Weight + Model\_year + I(\log(Weight))$

- Modelo intentando evitar el overfitting  $Mpg \sim Acceleration + I(\log(Weight)) + I(\log(Displacement))$

## 2.5. Ajustes con KNN

Sabemos que la función por defecto usa la distancia de Minkowski y escala los datos a igual rango. También usa un  $k$  de 7, y sería recomendable probar con varios.

Vamos a probar con diferentes modelos, primero el multivariable con todas

[1] 1.880835

Y probando con varios obtenemos el menor error con este

[1] 1.856269

Que visualmente nos quedaría

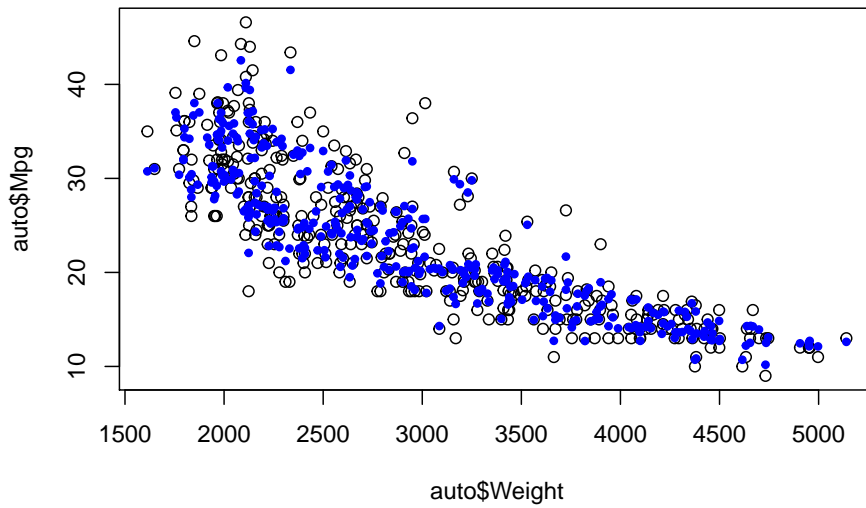


Figura 52

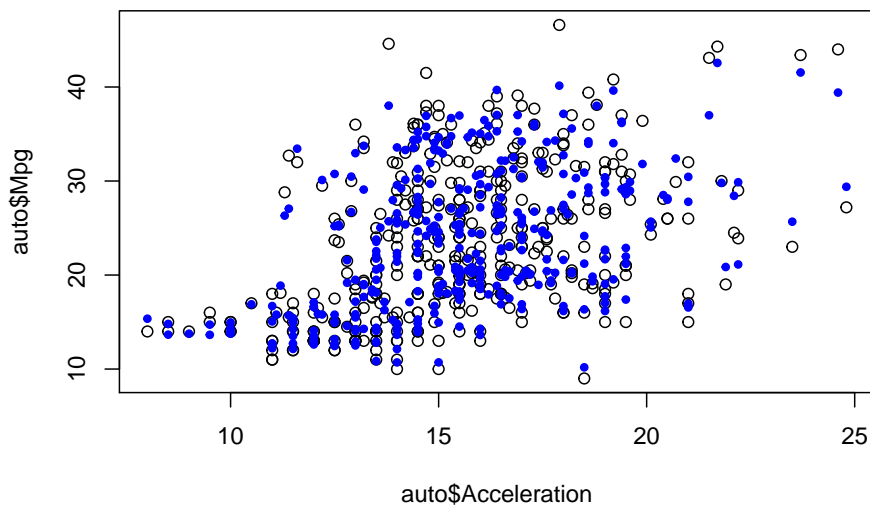


Figura 53

Si probamos el modelo no lineal obtenido en los pasos anteriores (con mejor R2)

[1] 2.104086

Nos da peor error.

Y si probamos el modelo no lineal en el que intemos resolver el overfitting

[1] 2.938051

Obtenemos aún mayor empeoramiento. Gráficamente:

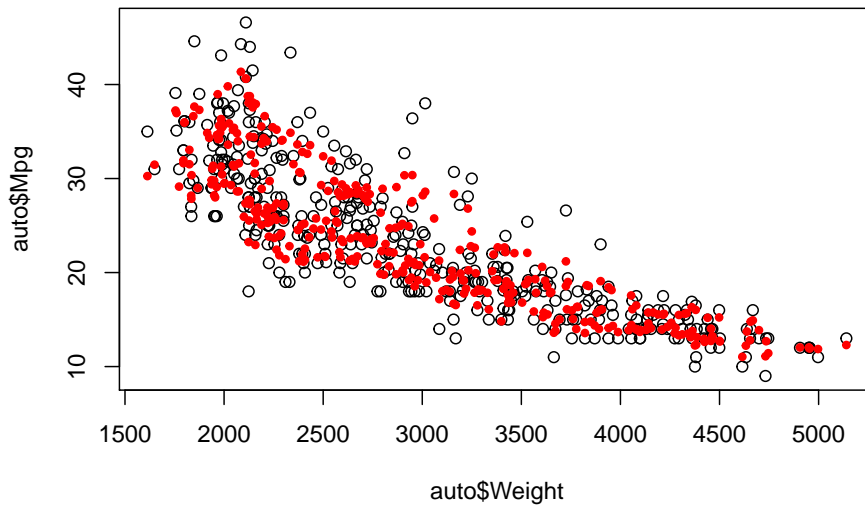


Figura 54

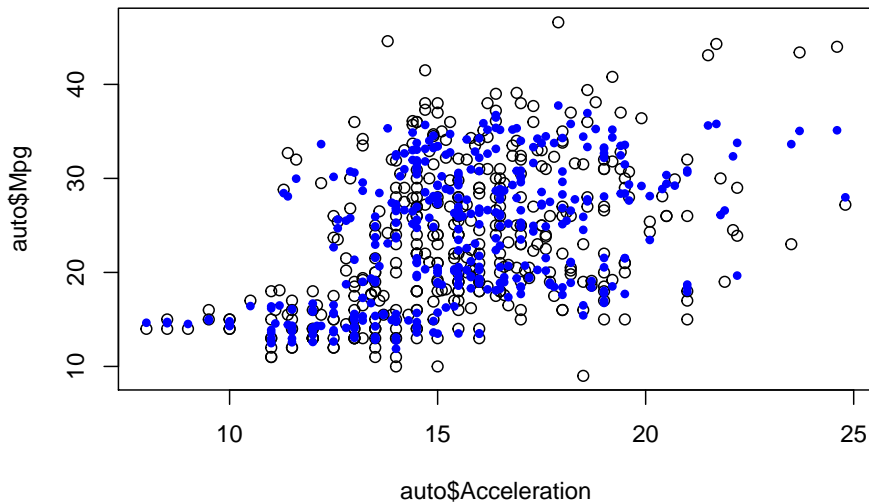


Figura 55

El RMSE (Root Mean Square Error) o raíz del error cuadrático medio nos permite calcular el error en nuestro conjunto de test o training producido en las predicciones, definido por la fórmula:

El método para evitar el overfitting que usamos en el apartado anterior probablemente no funcione con KNN por seguir una metodología totalmente diferente. El ajuste de KNN para regresión no tiene nada que ver con los modelos LM. Podemos aún así guardarlo para comprobarlo.

Tendríamos por tanto los siguientes modelos para KNN:  $Mpg \sim . - Acceleration$   $Mpg \sim Acceleration + I(\log(Weight)) + I(\log(Displacement))$

Comparandolos gráficamente, vemos que son similares

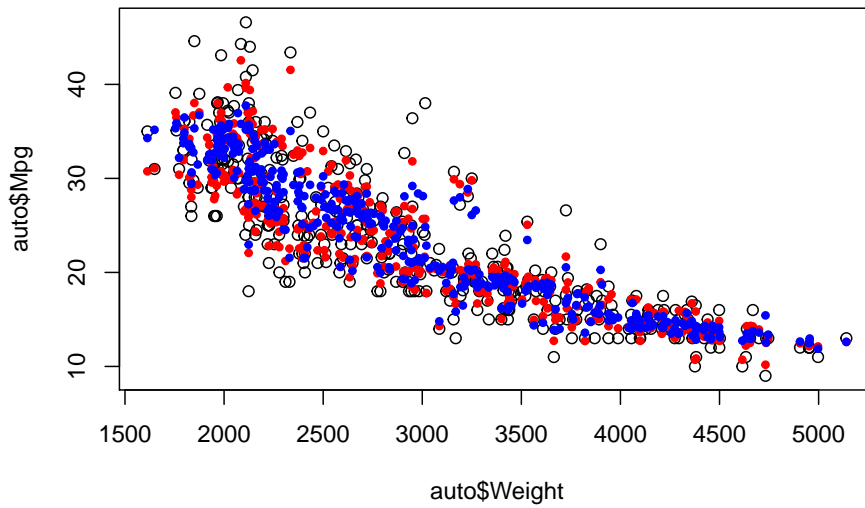


Figura 56

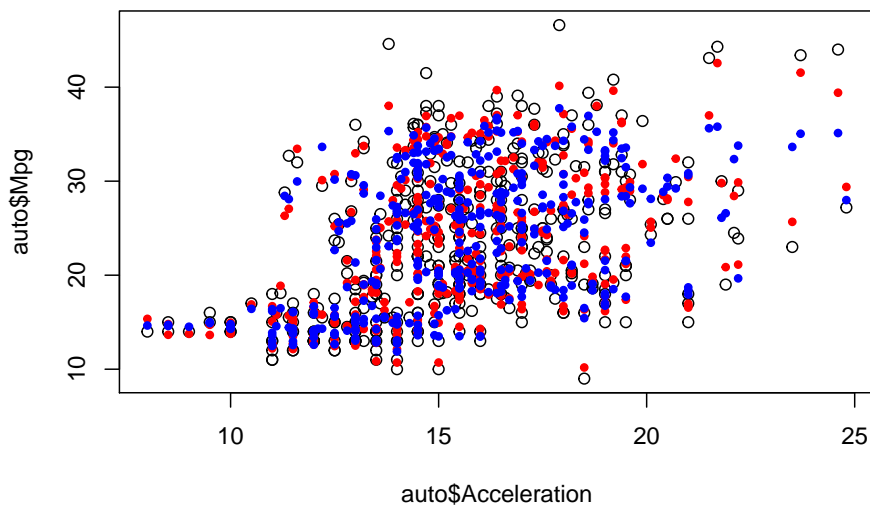


Figura 57

Aunque el que intenta evitar el overfitting (en color azul en la gráfica), tiene menor dispersión.

## 2.6. Comparativa de los ajustes anteriores con cross-validation

Recordamos los modelos obtenidos: LM:  $\text{Mpg} \sim \text{Weight} + \text{Model\_year} + \text{I}(\log(\text{Weight}))$   
 $\text{Mpg} \sim \text{Acceleration} + \text{I}(\log(\text{Weight})) + \text{I}(\log(\text{Displacement}))$   
 KNN:  $\text{Mpg} \sim . - \text{Acceleration}$   $\text{Mpg} \sim \text{Acceleration} + \text{I}(\log(\text{Weight})) + \text{I}(\log(\text{Displacement}))$   
 $\text{c}(\text{"Displacement"}, \text{"Horse\_power"}, \text{"Weight"}, \text{"Acceleration"}, \text{"Model\_year"}, \text{"Mpg"})$  X1  
 X2 X3 X4 X5 Y  
 LM:  $Y \sim X3 + X5 + \text{I}(\log(X3))$   $Y \sim X4 + \text{I}(\log(X3)) + \text{I}(\log(X1))$   
 KNN:  $Y \sim . - X4$   $Y \sim X4 + \text{I}(\log(X3)) + \text{I}(\log(X1))$   
 Resultados obtenidos

Regresión 1: [1] 9.333066

Regresión 2: [1] 17.10519

KNN 1: [1] 7.291517

KNN 2: [1] 18.43846

Con el proceso de cross-validation dividimos el dataset en N subconjuntos (folds) y repetimos el entrenamiento N veces. Cada entrenamiento se aplica reservando uno de los subconjuntos como test y entrenando con el resto. Al final, el error obtenido para el modelo es la media de los errores en cada fold. La elección del número de folds es importante y si el problema lo permite (en términos de gasto computacional), se debería probar con varios. En este caso hemos utilizado 5 folds.

Con esto conseguimos no desperdiciar el conocimiento del conjunto de test y no guiarnos por una única evaluación del modelo.

Como resultados, nos muestra que los modelos con los que obtuvimos mejores resultados de R2 y RSME en sus apartados han acabado con mejor RSME tras el cross-validation. También apreciamos que con KNN conseguimos ligeramente mejores resultados.

Por completitud, mostramos también los resultados en training

Regresión 1: [1] 9.159891

Regresión 2: [1] 16.62285

KNN 1: [1] 3.659828

KNN 2: [1] 8.642349

Que nos muestran que ninguno de los modelos LM estaban haciendo overfitting, pero en cambio en KNN si existe una diferencia significativa entre training y test.

## 2.7. Comparativa de tests

Para comparar los algoritmos vamos a aplicar test estadísticos en base a los resultados obtenidos en múltiples datasets. Para asegurar la igualdad de condiciones los algoritmos hacen uso de parámetros genéricos y utilizan las mismas particiones de cross-validation.

Estas son las tablas de resultados que tenemos:

out_test_lm	out_test_kknn
0.1909091	0.1000000
0.1000000	1.0294118
0.1000000	0.4339071
0.1000000	0.3885965
0.1548506	0.1000000
0.1000000	0.3061057



Aplicamos el test de Wilcoxon a LM y KNN

```
V
78
V
93
p-value: [1] 0.7660294
```

Obtenemos un ranking de 78 para LM y 93 para KNN, con un p-valor de 0.77 (o nivel de confianza del 33%).

Esto nos dice que gana KNN pero puesto que el p-value no es lo suficientemente grande no podemos afirmar con un nivel alto de significación que las diferencias entre los tests sean notorias.

Ahora aplicamos en test de Friedman a los dos algoritmos anteriores junto al algoritmo M5:

```
Friedman rank sum test

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

El p-value es  $<0.05$  por lo que podemos concluir que al menos hay un par de algoritmos de calidad diferente.

Vemos cuáles de ellos lo son haciendo el test post-hoc de HOLM

```
Pairwise comparisons using Wilcoxon signed rank exact test

data: as.matrix(tablatst) and groups

 1      2
2 0.580 -
3 0.081 0.108
```

```
P value adjustment method: holm
```

Con el test post-hoc de HOLM podemos asegurar que 3-1 (M5 vs LM) son diferentes. También podemos afirmar M5 respecto de KNN pero con un nivel de confianza menor.

De KNN y LM no podemos afirmar nada puesto que el p-valor es extremadamente grande.

### **3. Clasificación: Análisis Estadístico de Datos**

## 4. Técnicas de Clasificación

## Referencias

[1] <http://lib.stat.cmu.edu/datasets/cars.desc>.