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

# 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, en base a los cinco atributos, el consumo de Mpg de un nuevo coche:

6. **Mpg**: Millas-por-galón, indica la cantidad de galones (1G +-= 3,78L) de fuél que consume un vehículo al recorrer una milla (1m +-= 1,6km).

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 plantemos 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 plantearse esta cuestión 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 otro lado, no tenemos información más precisa para los meses dentro del año)

En un principio, el dataset está planteado para regresión, por lo que 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

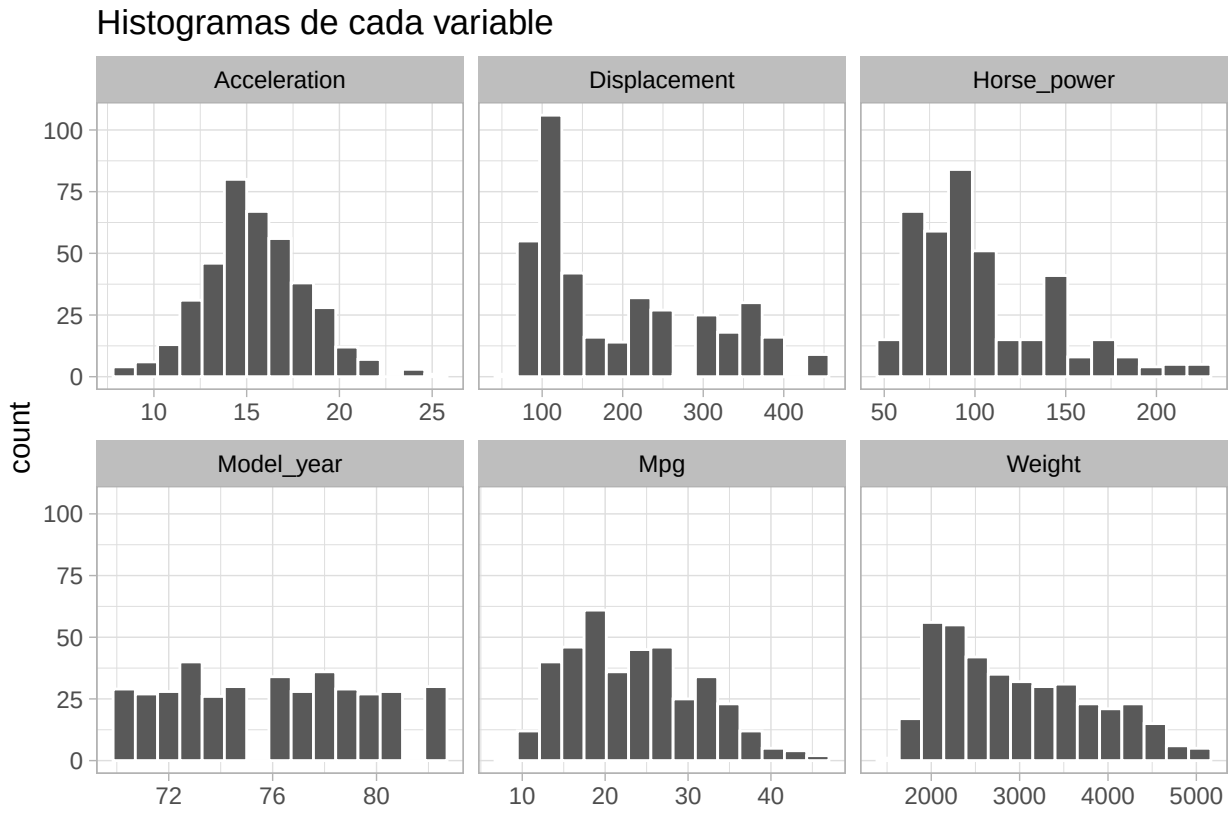
| Displacement | Horse_power | Weight | Acceleration | Model_year | Mpg |
|---|---|---|---|---|---|
| 91 | 70 | 1955 | 20.5 | 71 | 26.0 |
| 232 | 100 | 2789 | 15.0 | 73 | 18.0 |
| 350 | 145 | 4055 | 12.0 | 76 | 13.0 |
| 318 | 140 | 4080 | 13.7 | 78 | 17.5 |
| 113 | 95 | 2372 | 15.0 | 70 | 24.0 |
| 97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

Hacemos summary para sacar datos de relevancia

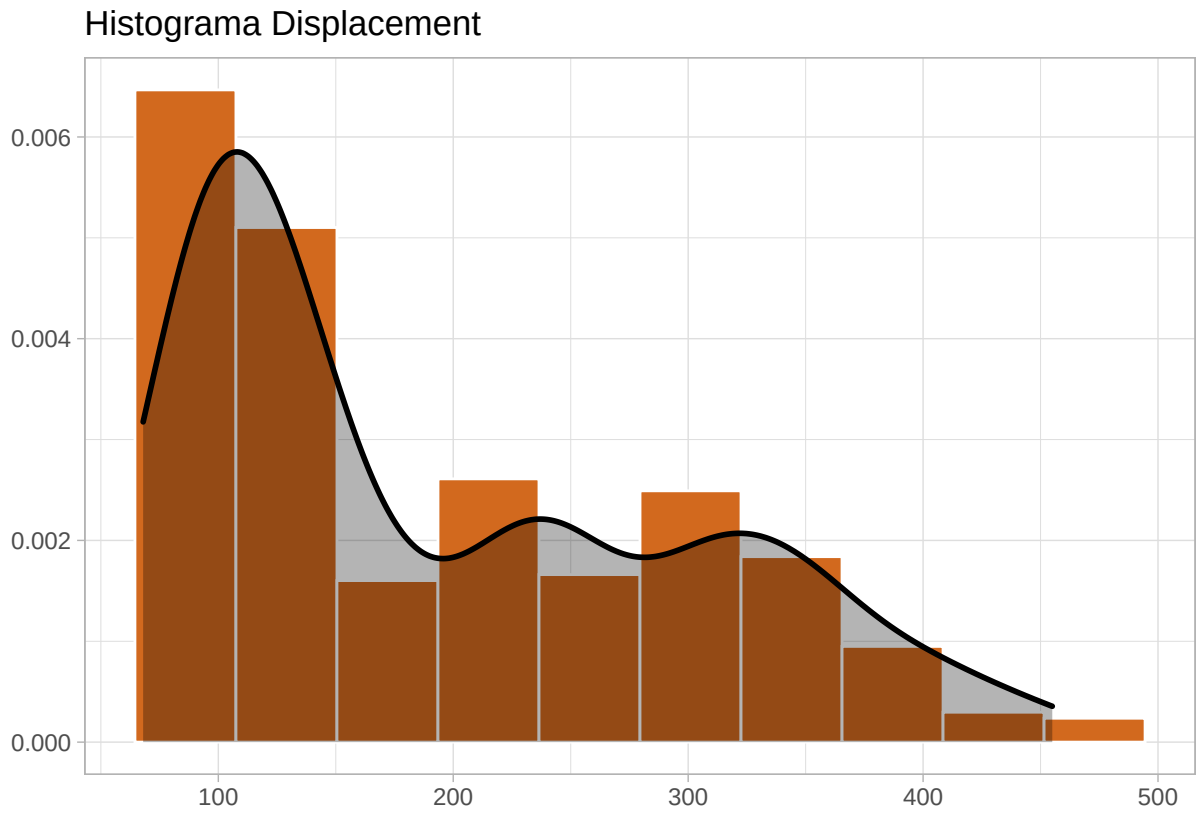
```
  Displacement     Horse_power        Weight      Acceleration     Model_year
 Min.   : 68.0   Min.   : 46.0   Min.   :1613   Min.   : 8.00   Min.   :70.00
```

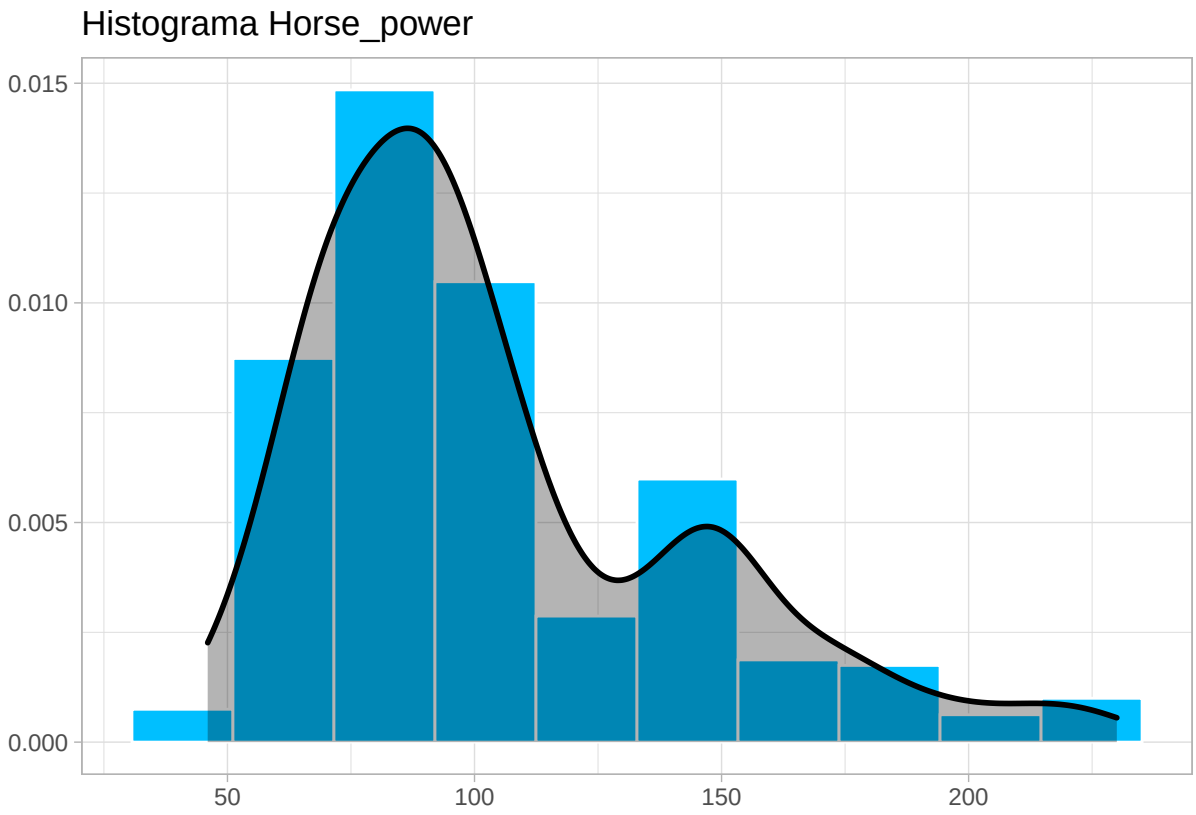
```
1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00
Median :151.0   Median : 93.5   Median :2804   Median :15.50   Median :76.00
Mean   :194.4   Mean   :104.5   Mean   :2978   Mean   :15.54   Mean   :75.98
3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615   3rd Qu.:17.02   3rd 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
```

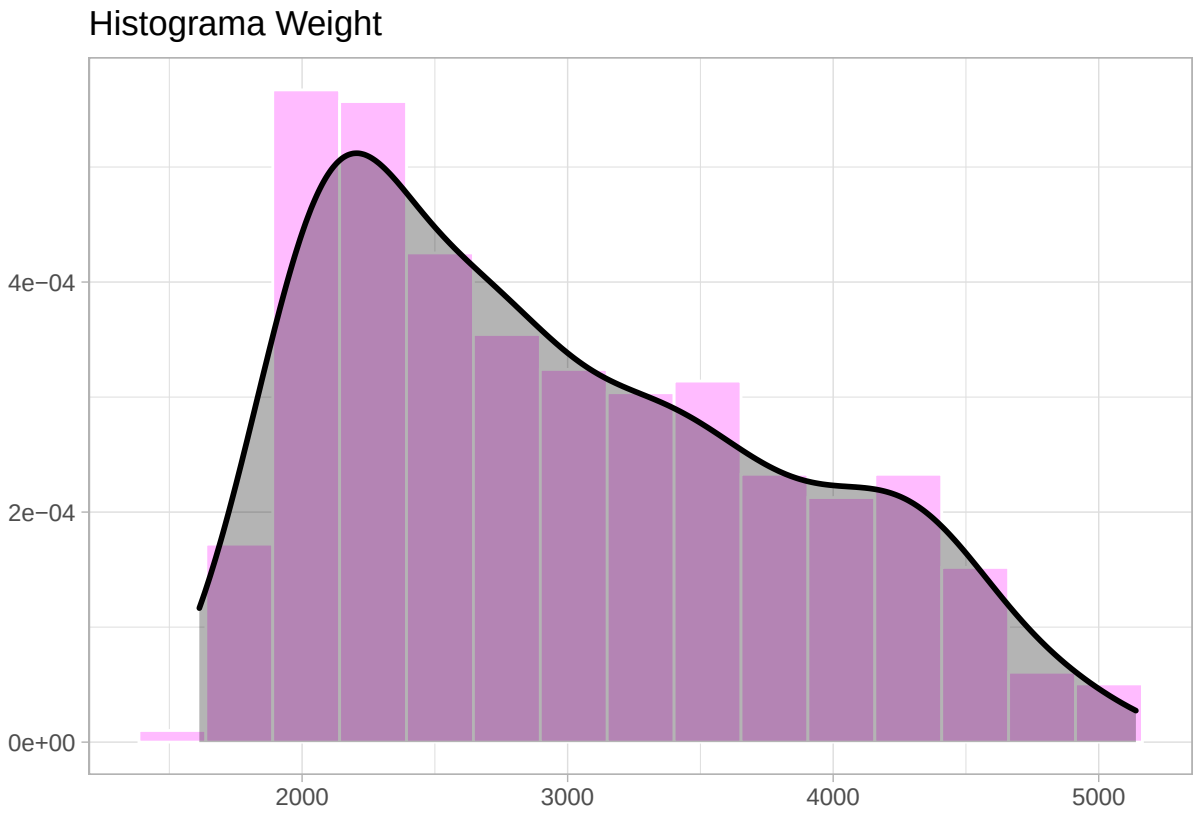
El dataset no cuenta con valores repetidos ni missing values.
Vamos a sacar plots de cada variable para verlo mejor

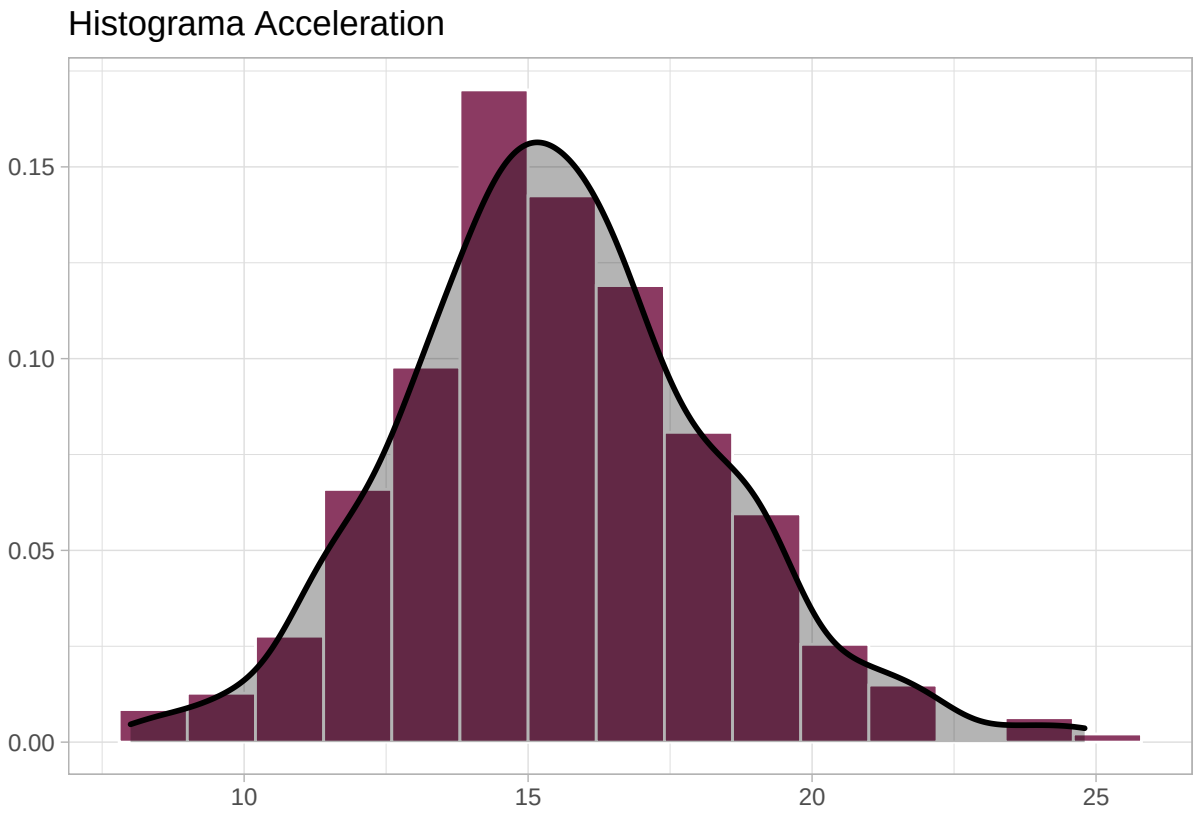


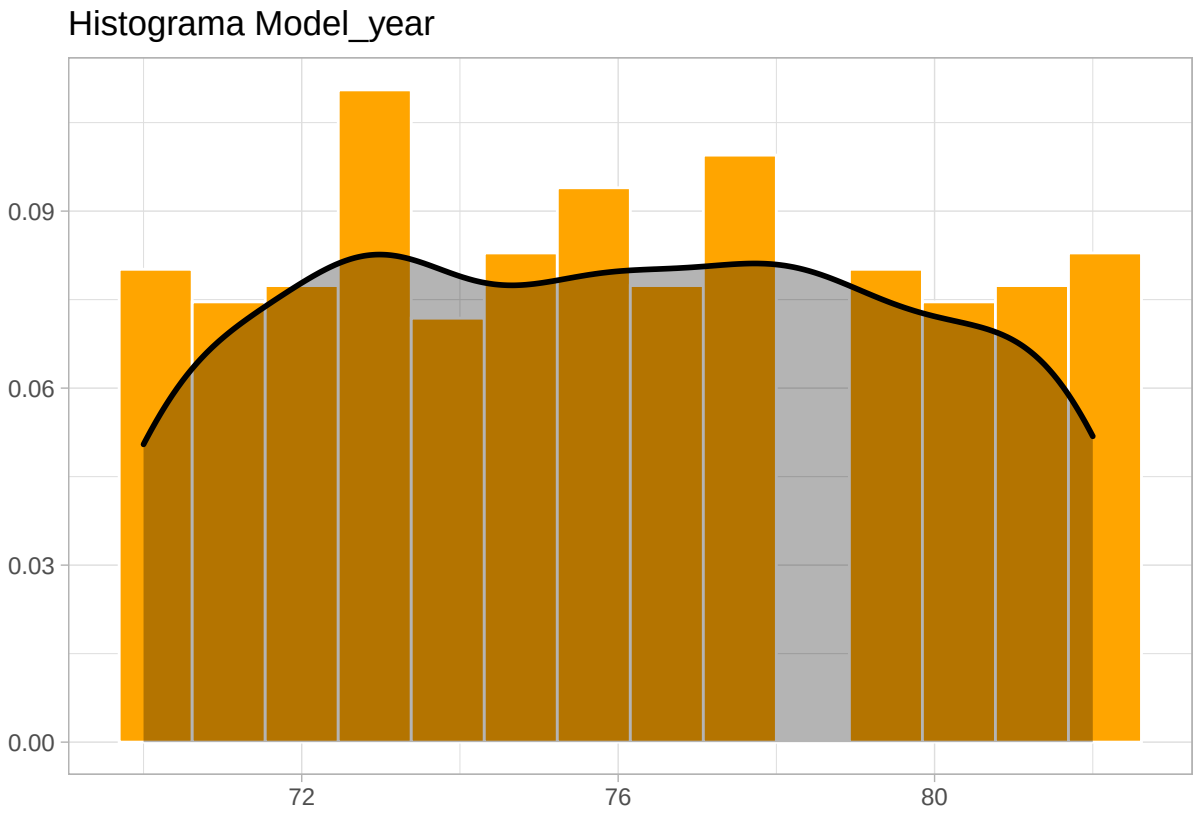

Una a una

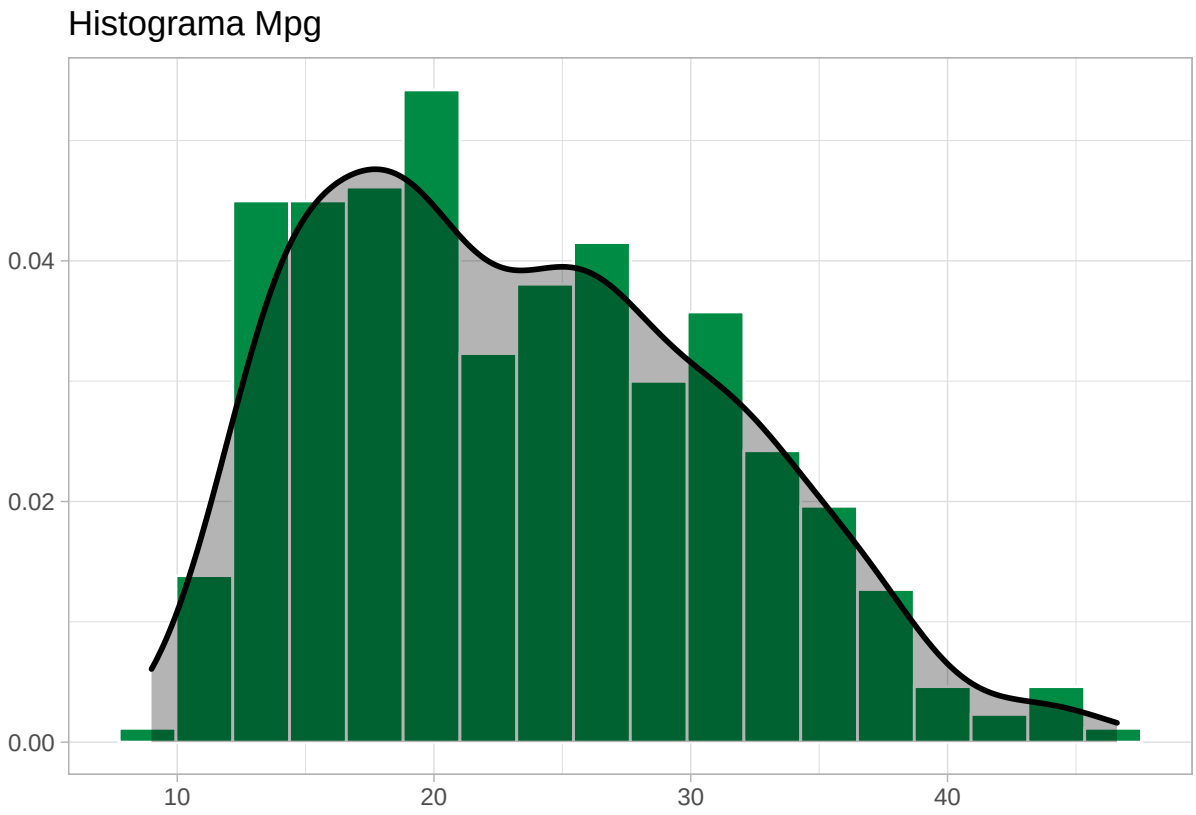

Histograma Displacement
0.006
0.004
0.002
0.000
100
200
300
400
500

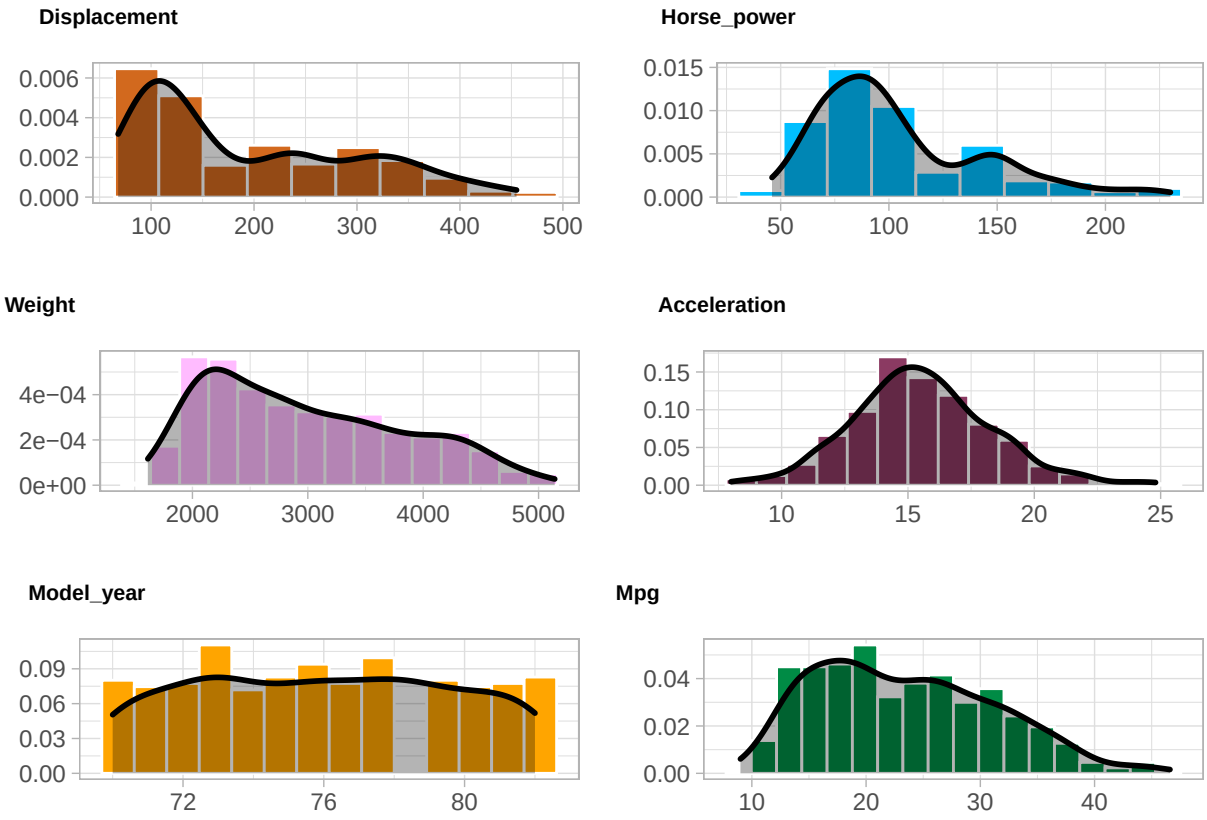

Histograma Horse_power
0.015
0.010
0.005
0.000
50
100
150
200


Histograma Weight
4e−04
2e−04
0e+00
2000
3000
4000
5000

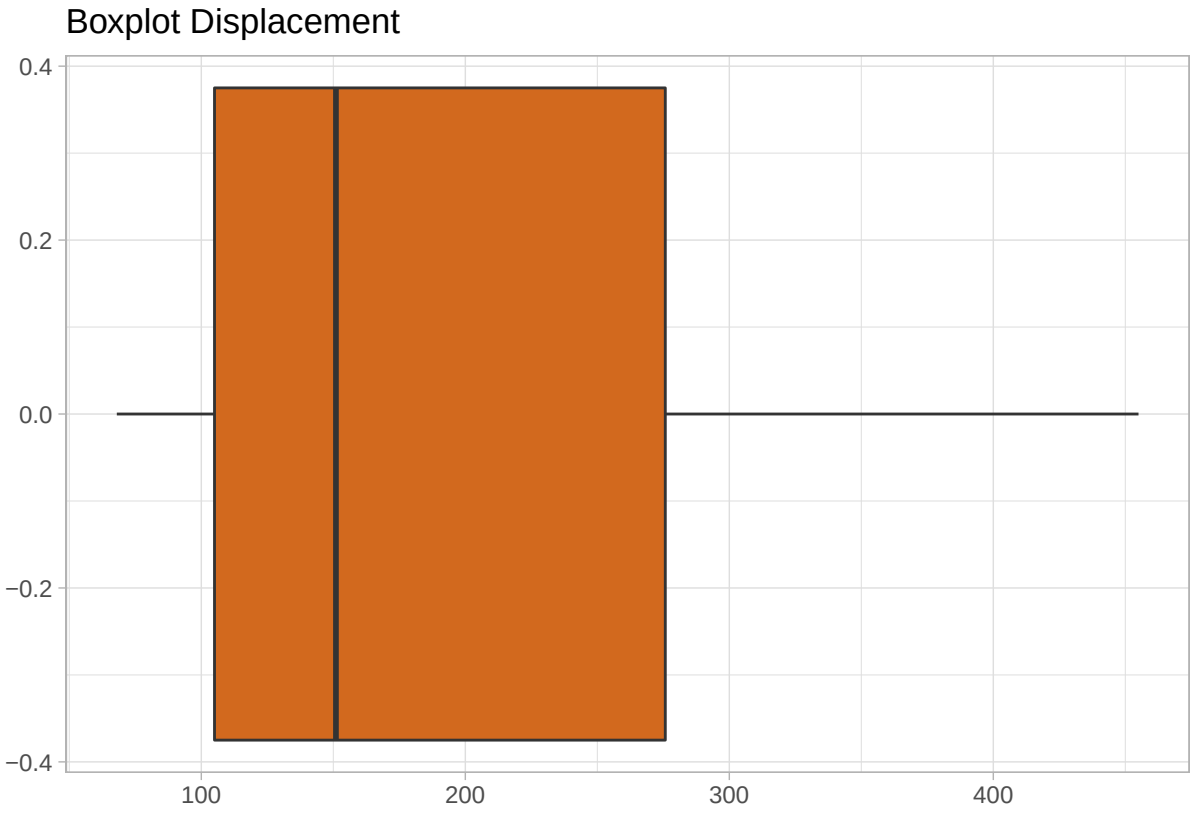

Histograma Acceleration
0.15
0.10
0.05
0.00
10
15
20
25

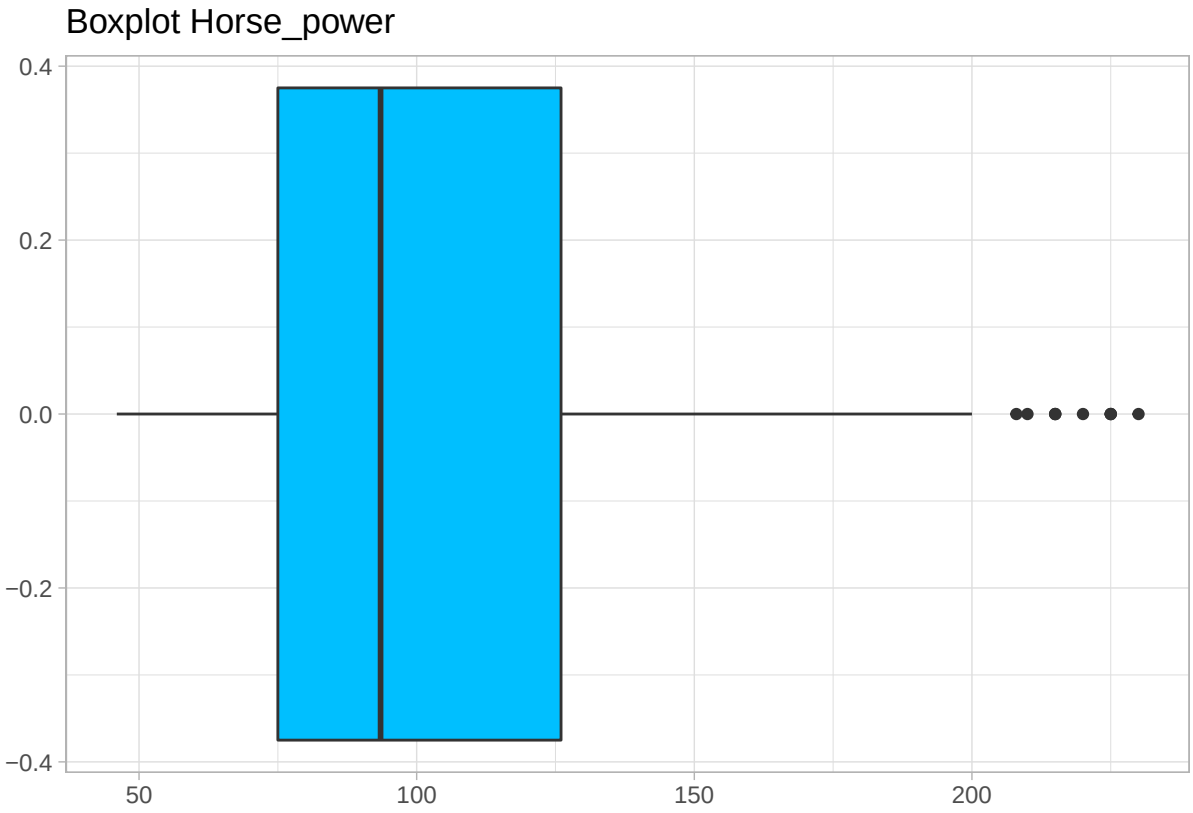

Histograma Model_year
0.09
0.06
0.03
0.00
72
76
80

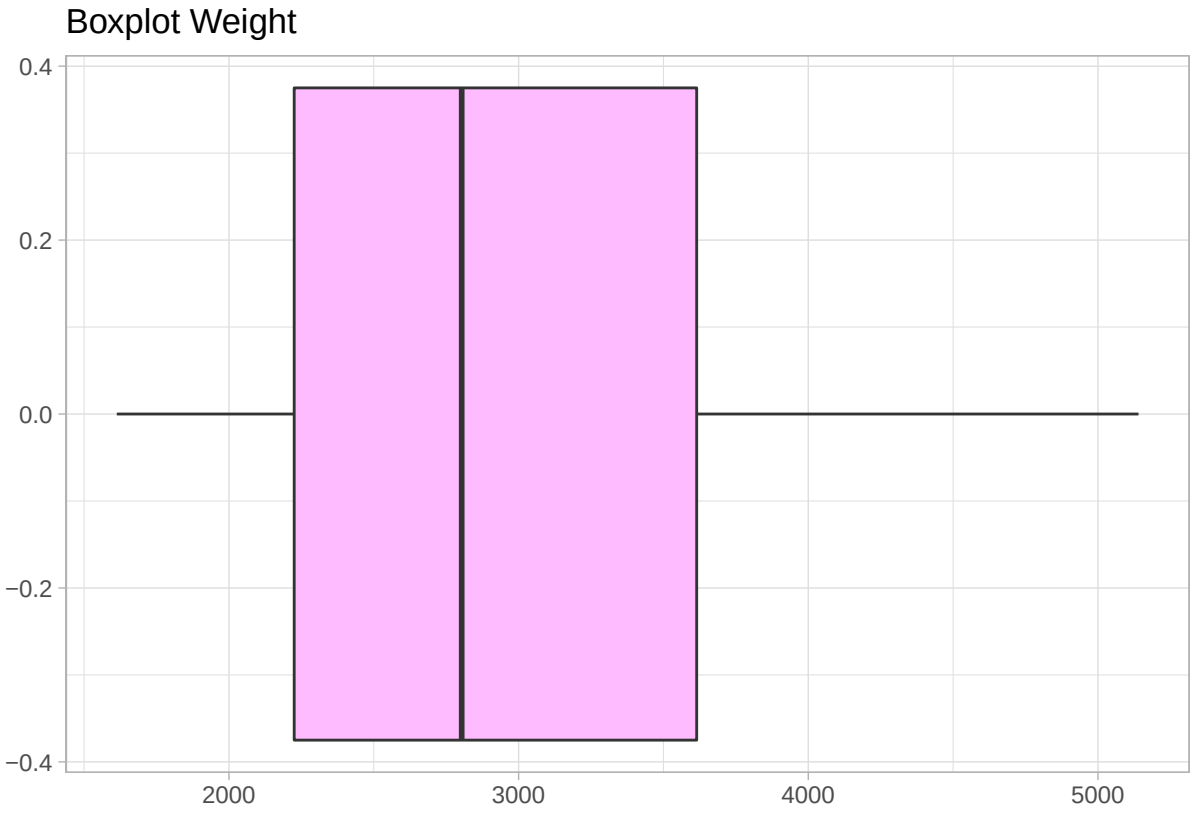

Histograma Mpg
0.04
0.02
0.00
10
20
30
40

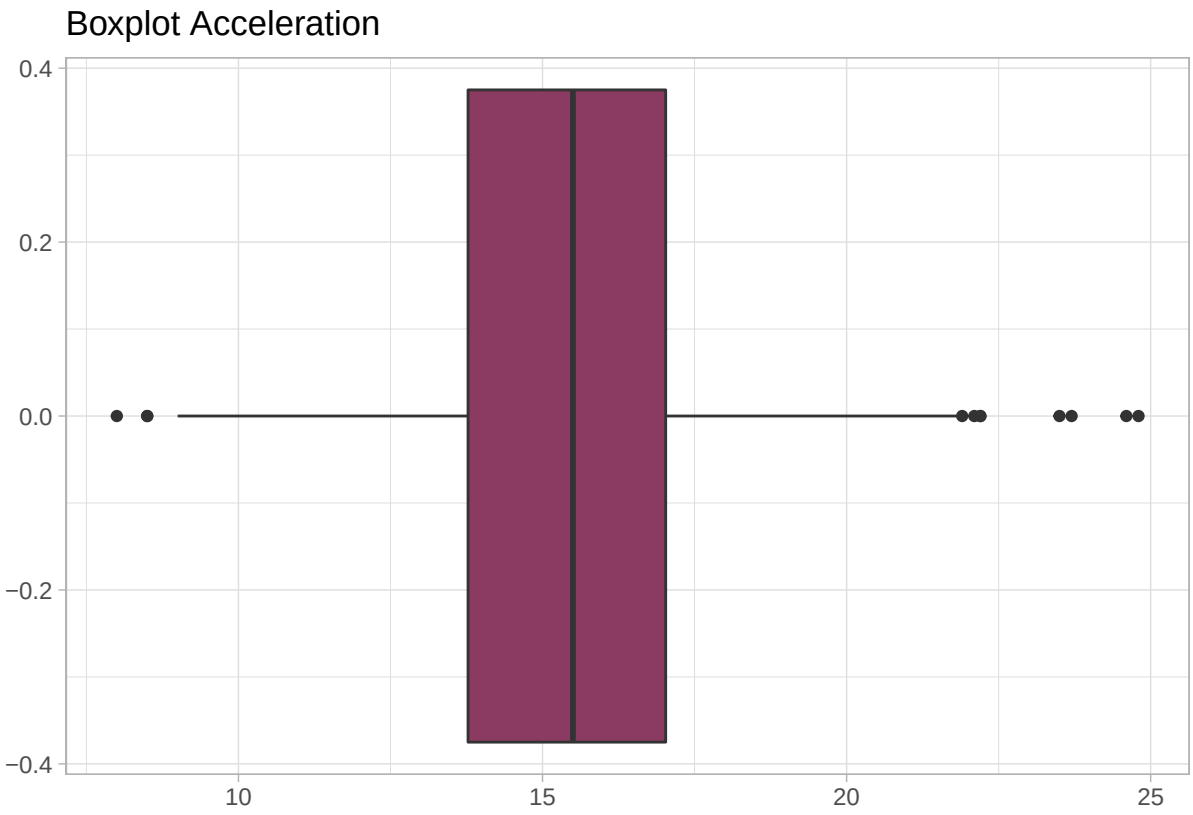


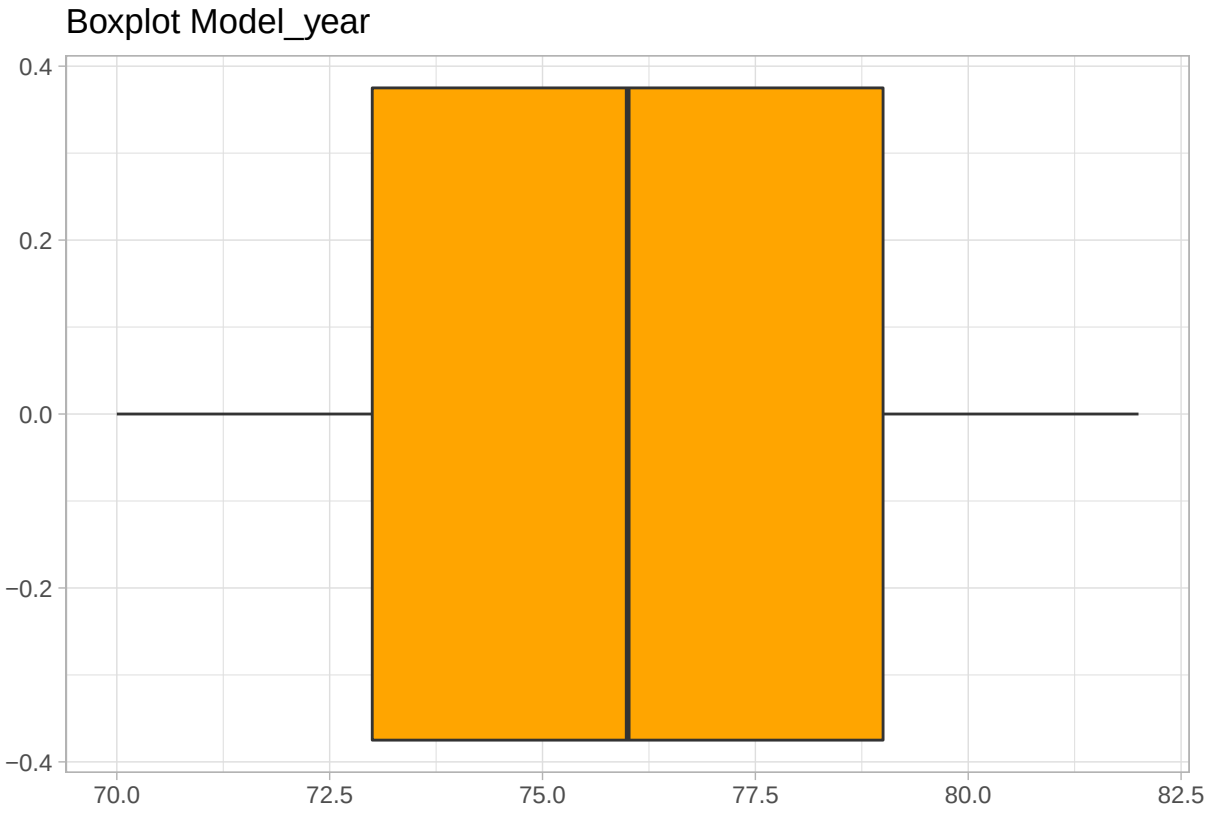

Sobre la distribuciones de los datos

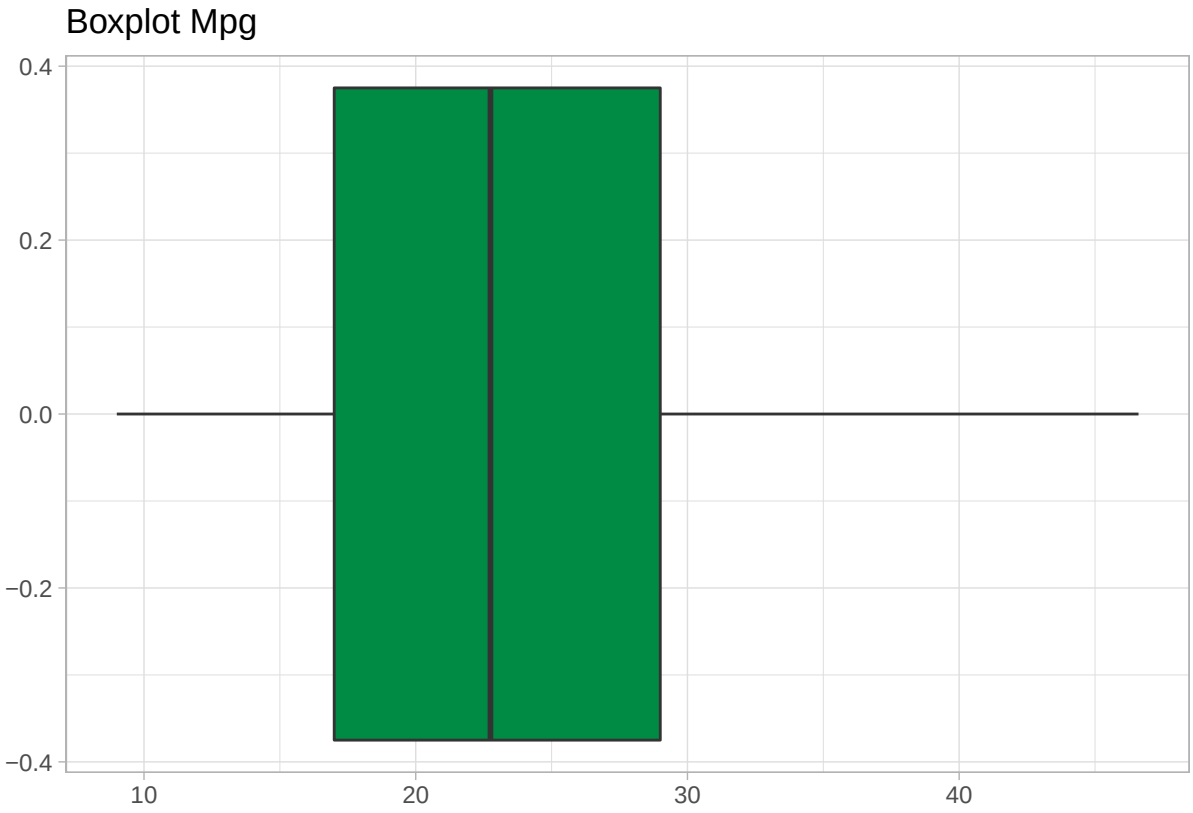

Boxplot Displacement
0.4
0.2
0.0
−0.2
−0.4
100
200
300
400

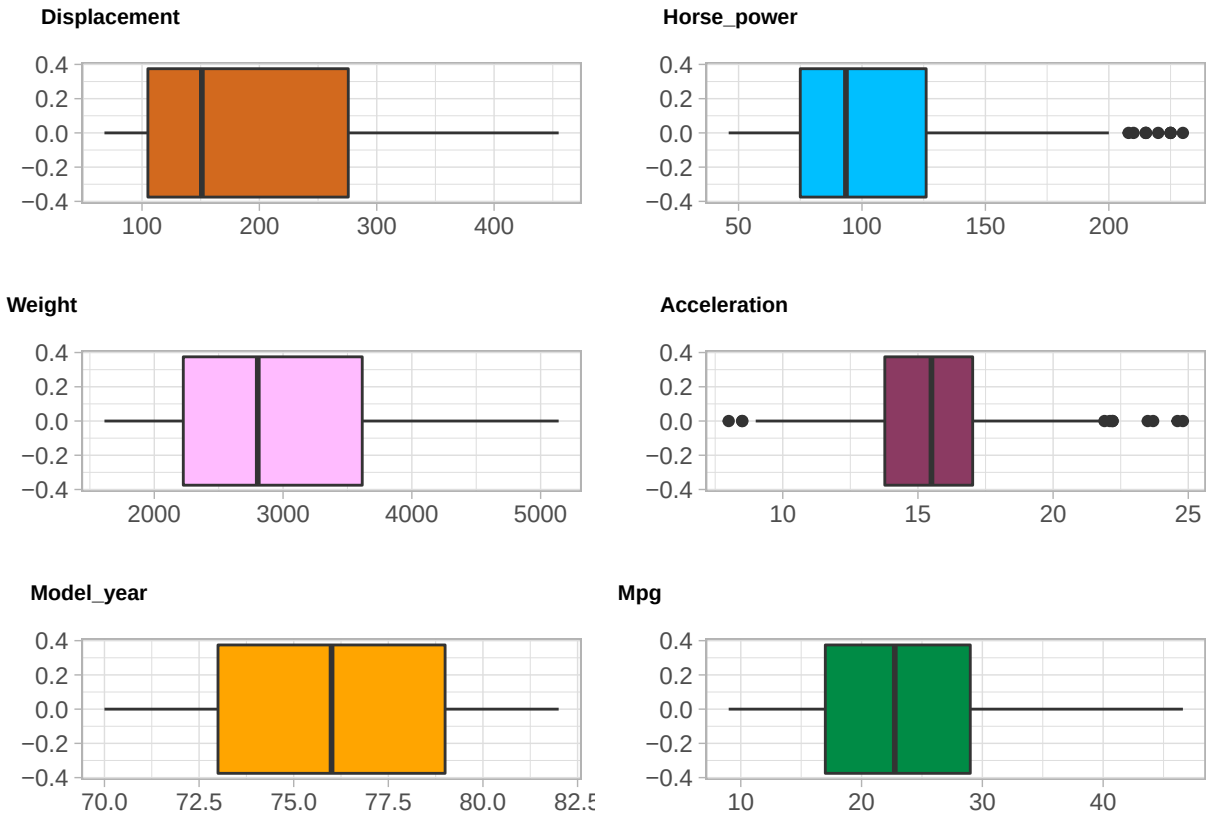

Boxplot Horse_power
0.4
0.2
0.0
−0.2
−0.4
50
100
150
200

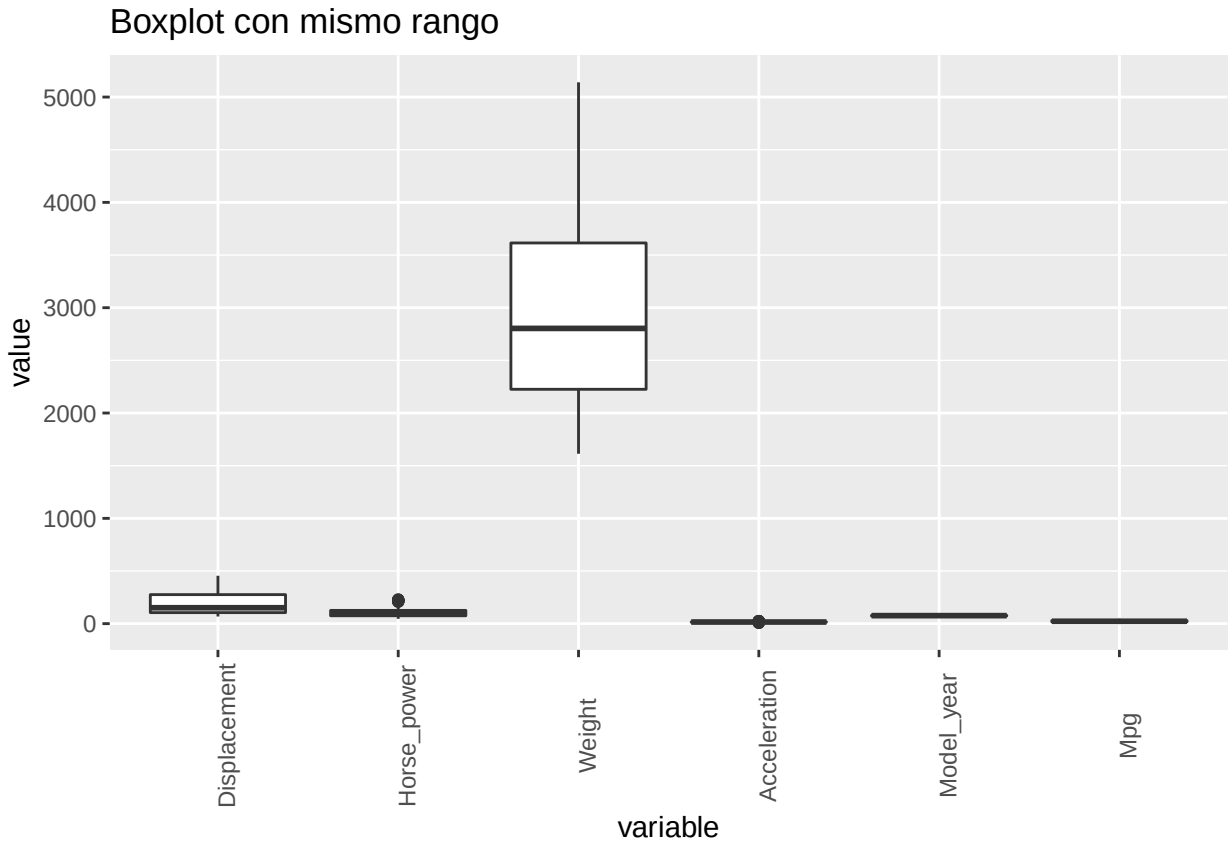

Boxplot Weight
0.4
0.2
0.0
−0.2
−0.4
2000
3000
4000
5000

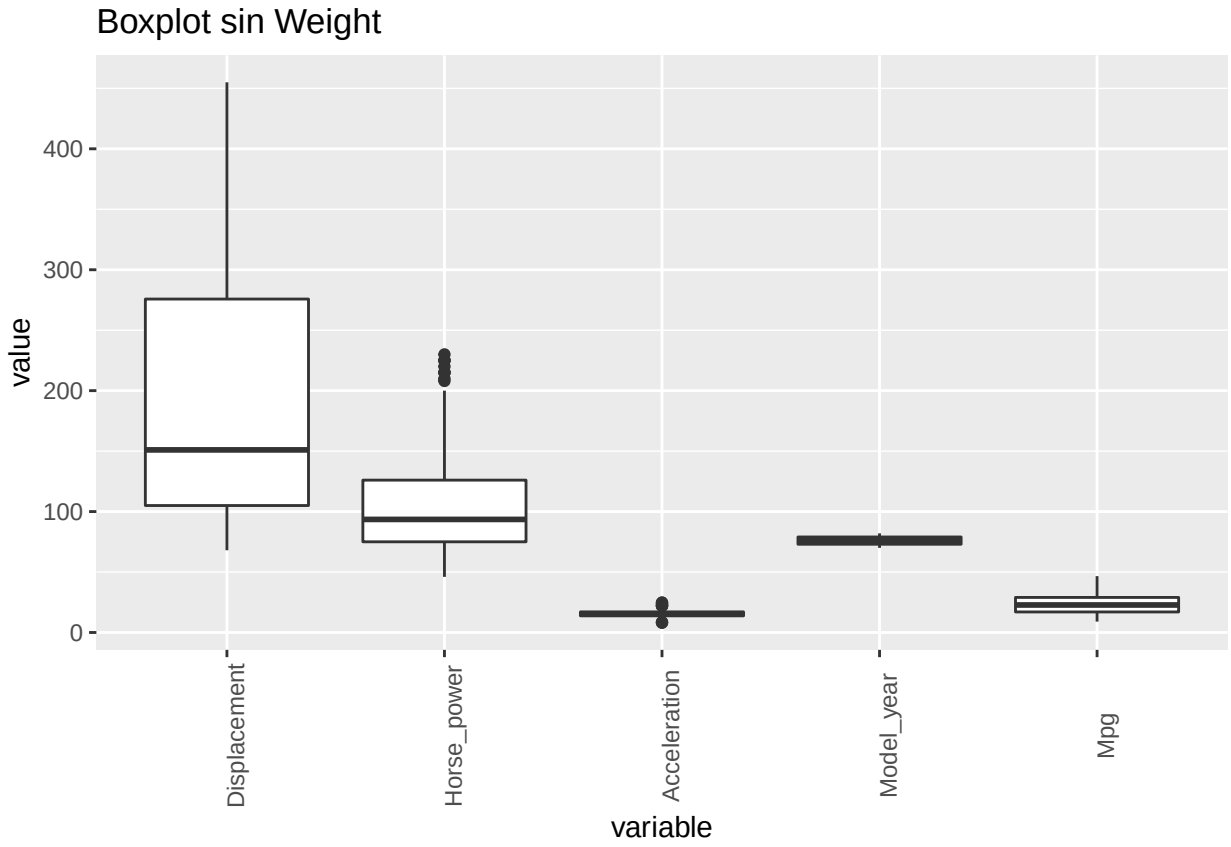

Boxplot Acceleration
0.4
0.2
0.0
−0.2
−0.4
10
15
20
25


Boxplot Model_year
0.4
0.2
0.0
−0.2
−0.4
70.0
72.5
75.0
77.5
80.0
82.5


Boxplot Mpg
0.4
0.2
0.0
−0.2
−0.4
10
20
30
40


Displacement
0.4
0.2
0.0
−0.2
−0.4
100
200
300
400
Horse_power
0.4
0.2
0.0
−0.2
−0.4
50
100
150
200
Weight
0.4
0.2
0.0
−0.2
−0.4
2000
3000
4000
5000
Acceleration
0.4
0.2
0.0
−0.2
−0.4
10
15
20
25
Model_year
0.4
0.2
0.0
−0.2
−0.4
70.0
72.5
75.0
77.5
80.0
Mpg
0.4
0.2
0.0
−0.2
−0.4
10
20
30
40


Boxplot con mismo rango
5000
4000
3000
2000
1000
0
value
Displacement
Horse_power
Weight
Acceleration
Model_year
Mpg
variable




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 intercuartiles 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 muya 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. Parece seguir 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):

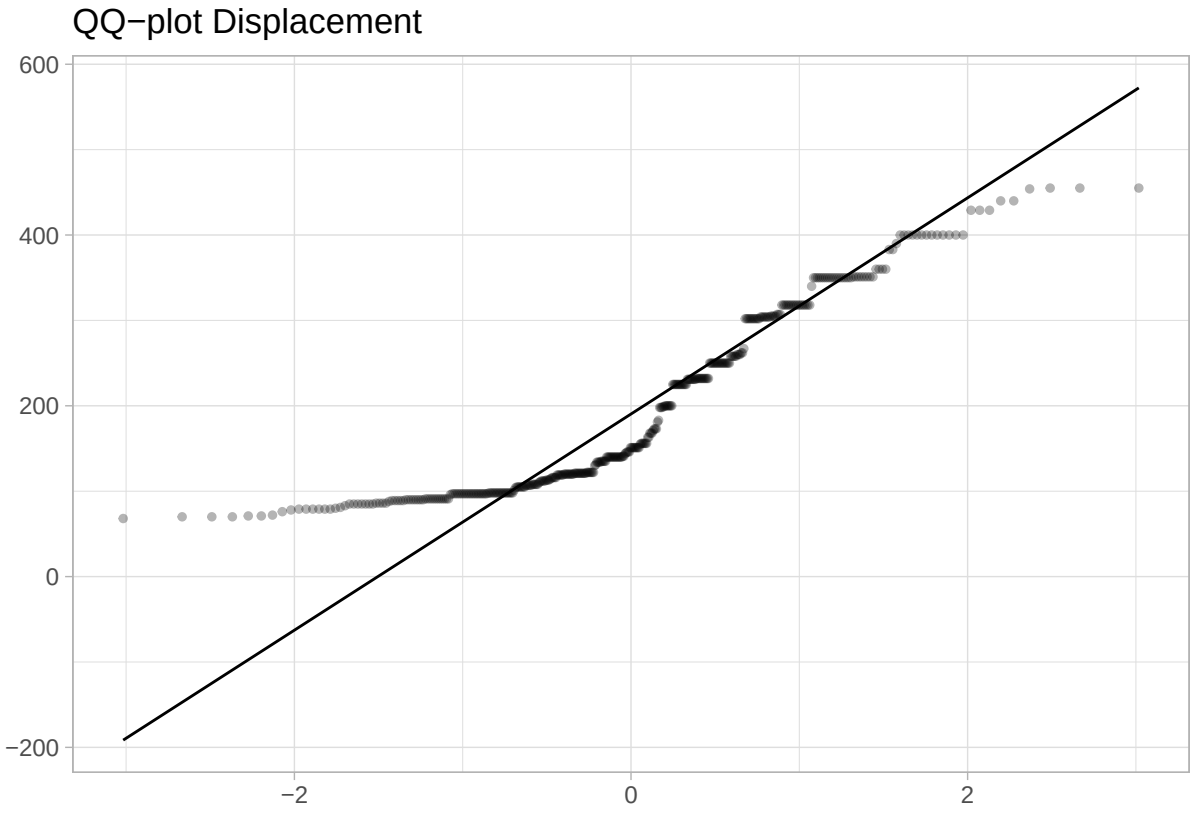
| vars | statistic | p_value | sample |
|---|---|---|---|
| Displacement | 0.8818359 | 0.0000000 | 392 |
| Horse_power | 0.9040975 | 0.0000000 | 392 |
| Weight | 0.9414661 | 0.0000000 | 392 |
| Acceleration | 0.9918671 | 0.0305289 | 392 |
| Model_year | 0.9469666 | 0.0000000 | 392 |
| Mpg | 0.9671696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

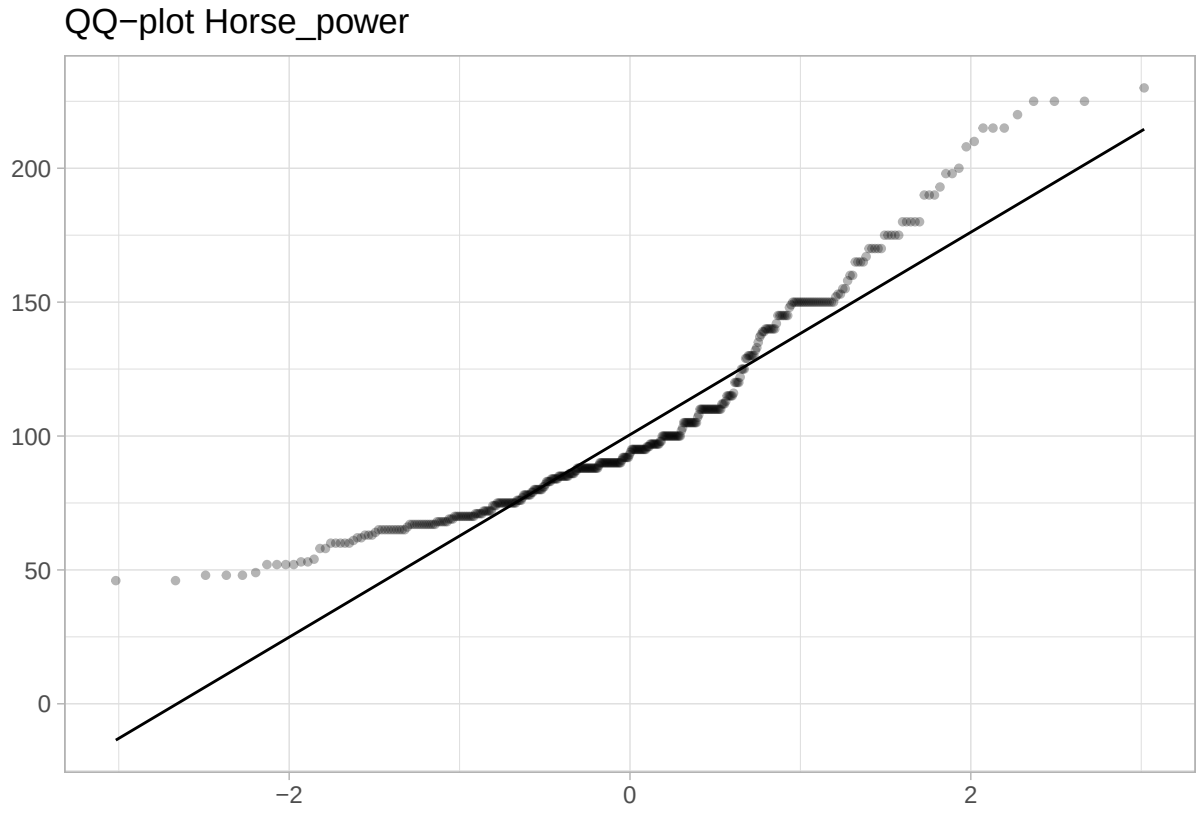
El test de Shapiro nos dice que ninguna variable sigue una distribución normal, con bastante certeza excepto en Acceleration.

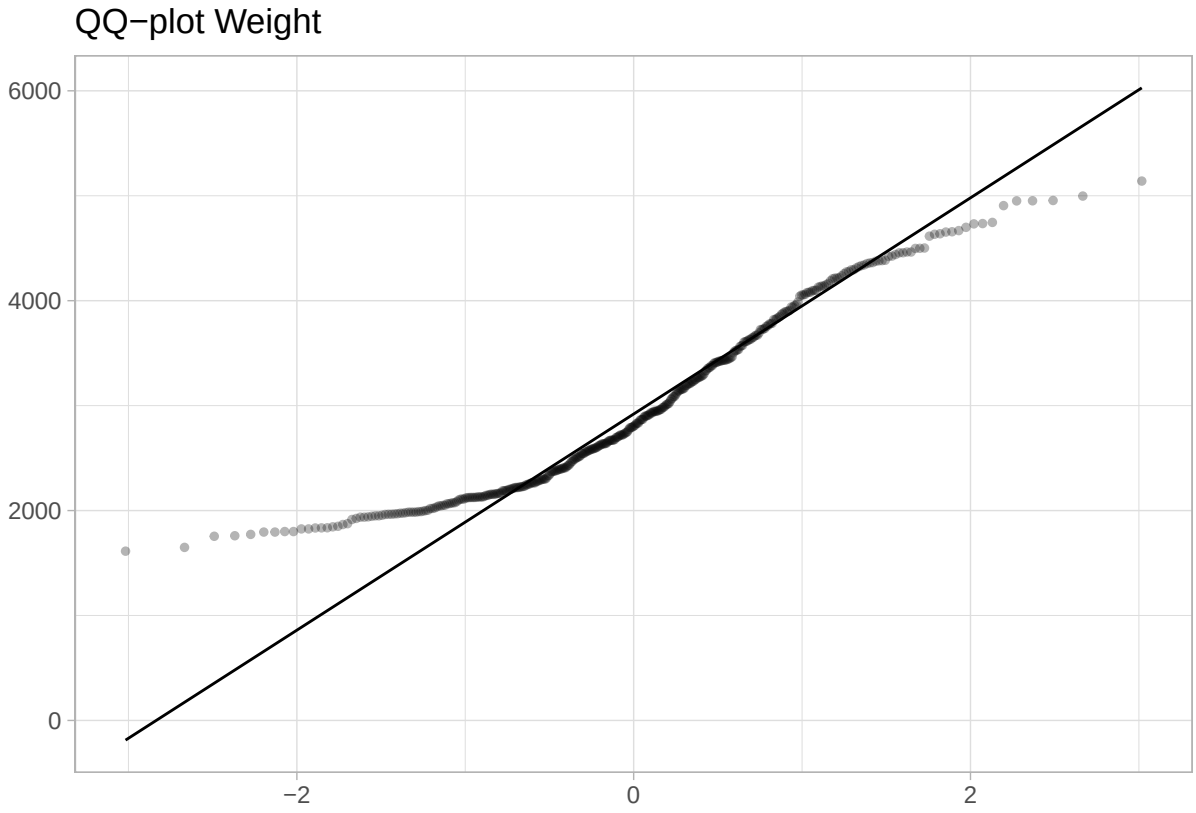
Para regresión aún así no es necesario.

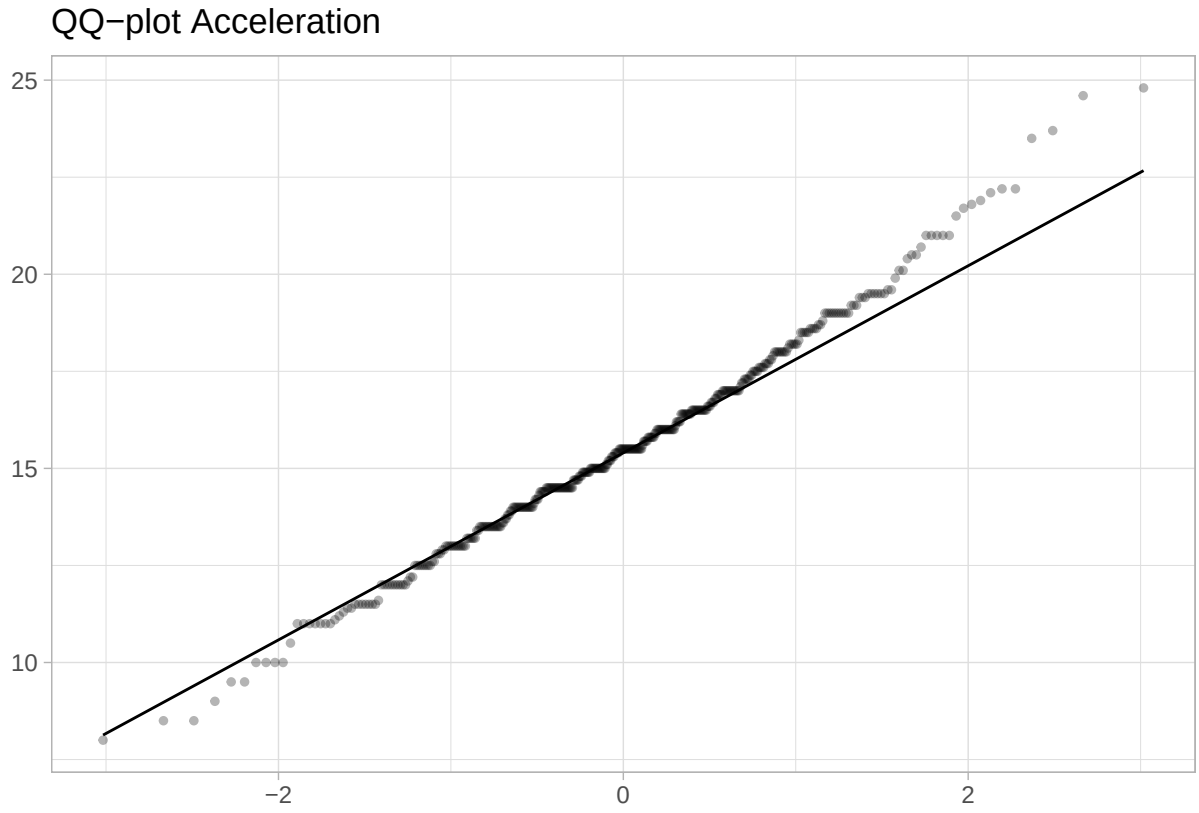
Se muestra aquí como no hay que dejarse engañar por los gráficos, puesto que Acceleration parecía seguirla. El p-value de Acceleration está muy cerca del umbral (0.03 vs 0.05). Es bastante probable de que la parte central derecha de la distribución sea la causante de no asegurar la normalidad.

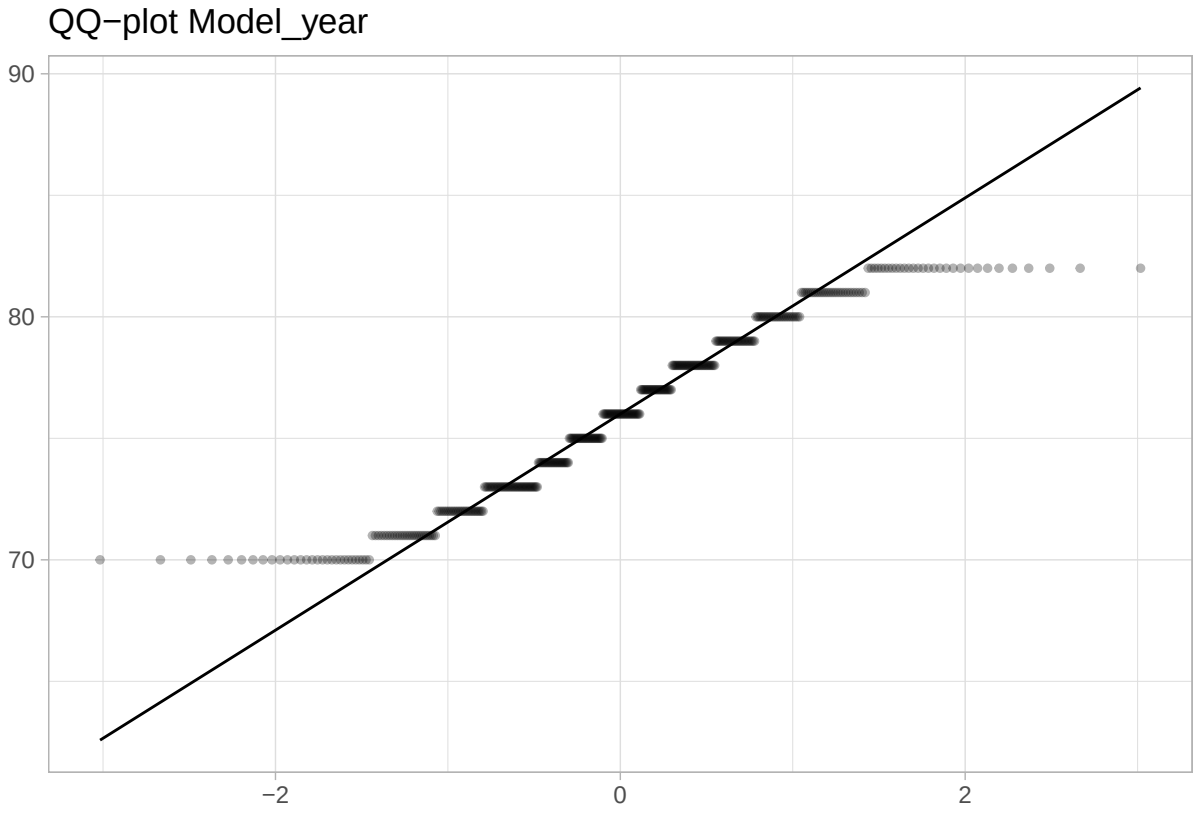
Vamos a mostrarlo con gráficos Q-Q para verlo mejor:

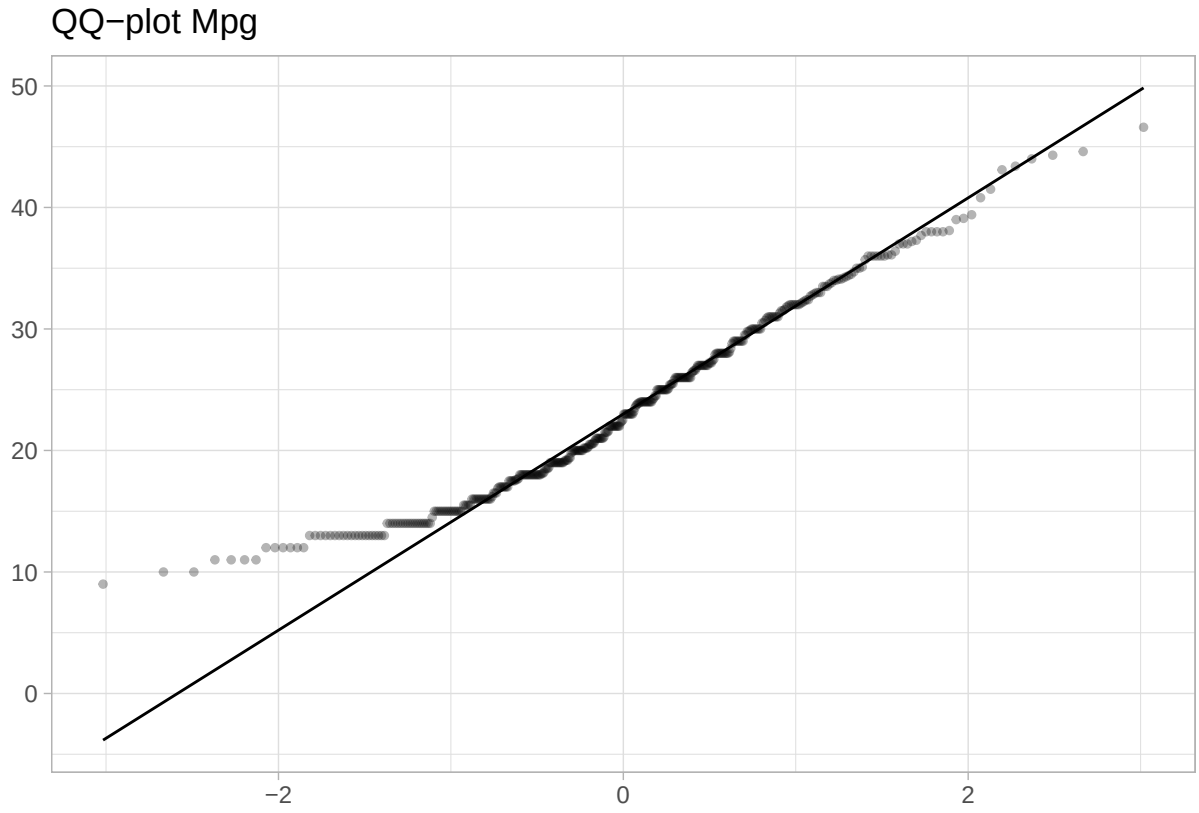

QQ−plot Displacement
600
400
200
0
−200
−2
0
2

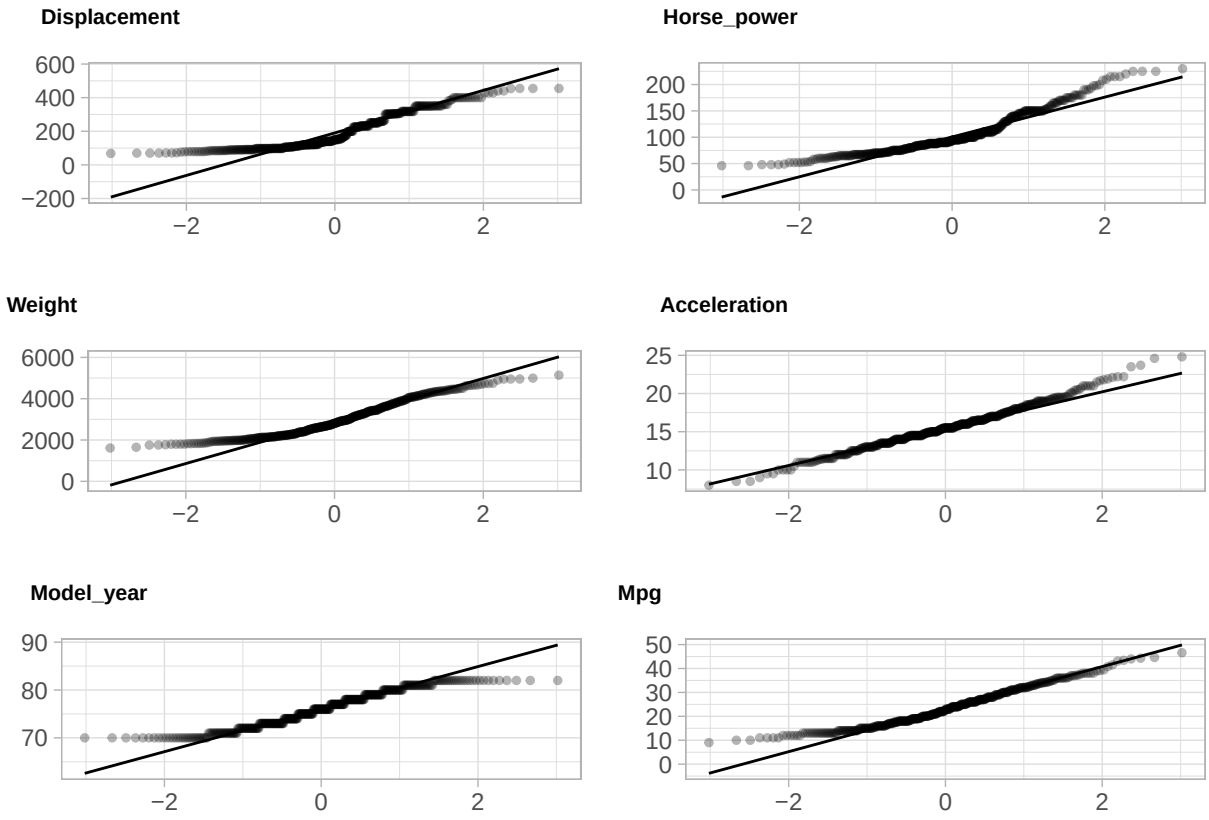

QQ−plot Horse_power
200
150
100
50
0
−2
0
2


QQ−plot Weight
6000
4000
2000
0
−2
0
2


QQ−plot Acceleration
25
20
15
10
−2
0
2


QQ−plot Model_year
90
80
70
−2
0
2


QQ−plot Mpg
50
40
30
20
10
0
−2
0
2




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.

# 2. Técnicas de Regresión

# 3. Clasificación: Análisis Estadístico de Datos

# 4. Técnicas de Clasificación

# Referencias

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