Practica 1 ApC

Ivan Peñarando Martínez
Joel Marco Quiroga Poma
Ferran Martínez Reyes

Apartado C

La base de datos

- Base de datos de coches (automobile.csv)
- 201 Filas, 26 Atributos → 9 Enteros, 10 Strings y 7 Floats
- Sin errores en el formato, Sin NaN's, Datos coherentes

Aun así, por cada atributo hemos comprobado:

- La media "atr.mean()" que sea coherente
- Valores únicos "atr.unique()" sin valores incorrectos
- Contar numero de elementos NaN de cada atributo→ Ninguno
- Significado del atributo i Tipo

```
symboling            0.841
normalized_losses  125.189
wheel_base          98.797
length             174.201
width               65.889
height              53.767
```

```
ENGINE_LOCATION

['front' 'rear']

WHEEL_BASE

[ 86.6  88.4  88.6  89.5  91.3  93.   93.1  93.3
  95.3  95.7  95.9  96.   96.1  96.3  96.5  96.6
```

```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   symboling          201 non-null    int64
 1   normalized_losses  201 non-null    int64
 2   make               201 non-null    object
 3   fuel_type          201 non-null    object
 4   aspiration         201 non-null    object
```

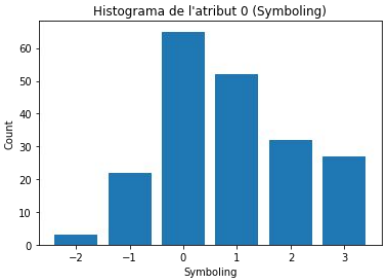
Atributos con distribución gausiana

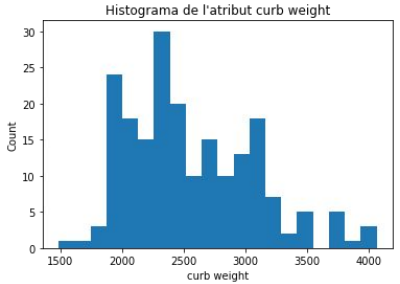
Descartamos: Todos los atributos discretos, o atributos no-numericos

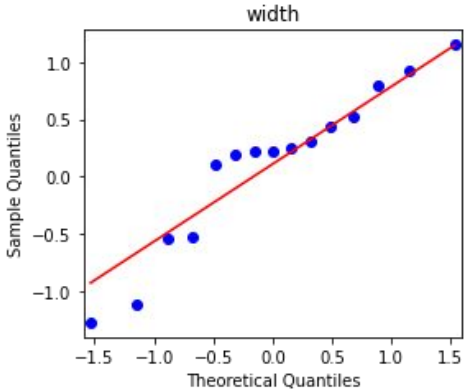
Tenemos en cuenta: Atributos continuos

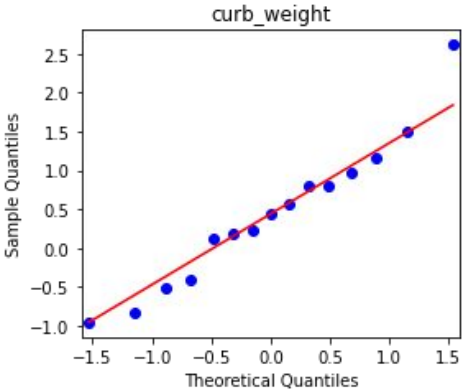
Test de normalidad:

- Histogramas poco fiables
- Método Q-Q Plot: Comparar distribuciones
  - A partir de dibujo de quartiles en una recta
  - Comparamos distribuciones con normales






width
Sample Quantiles
Theoretical Quantiles
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

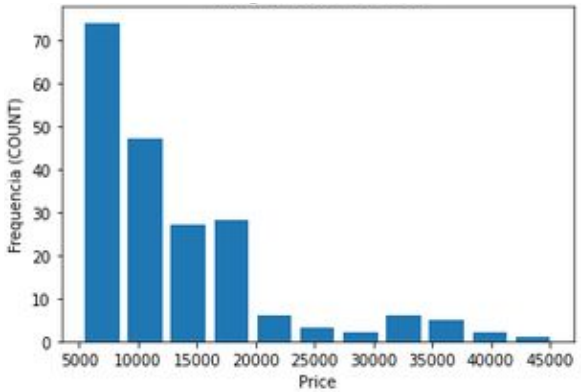

curb_weight
Sample Quantiles
Theoretical Quantiles
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Atributo Objetivo

El atributo PRICE

Razones:

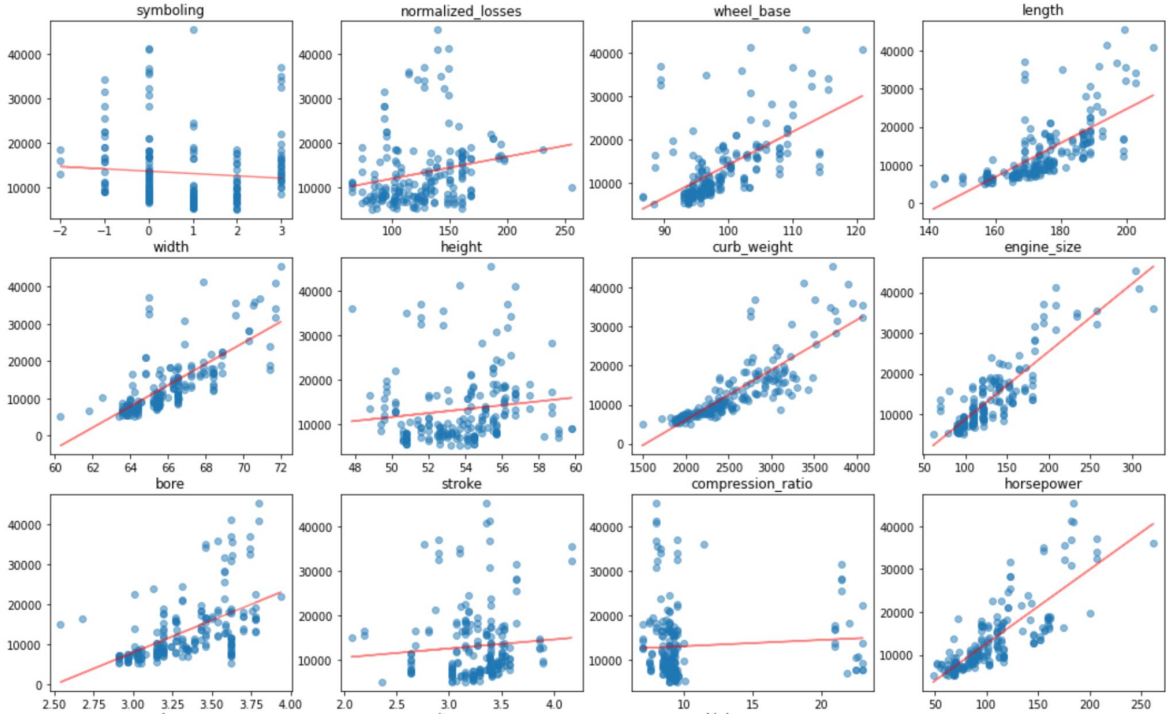
- Interés: Predecir el coste de un coche (Tasarlo de manera justa)
- Atributo lo suficientemente variable
- Atributo expresable en valores ordinales:
  - Precio Bajo (5000 - 15 000)
  - Precio Medio (15 000 - 30 000)
  - Precio Alto (30 000 - 45 000)



Apartado B

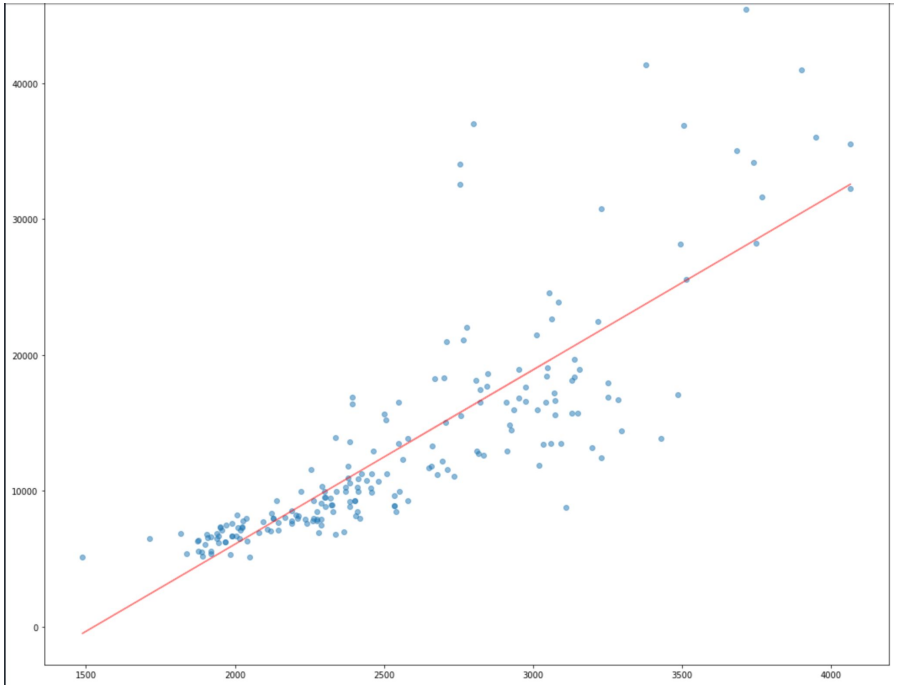
Primeras Regresiones

Realizamos regresiones para cada atributo con los datos sin normalizar para visualizar cuáles atributos aportan al regresor.



Primeres Regresiones

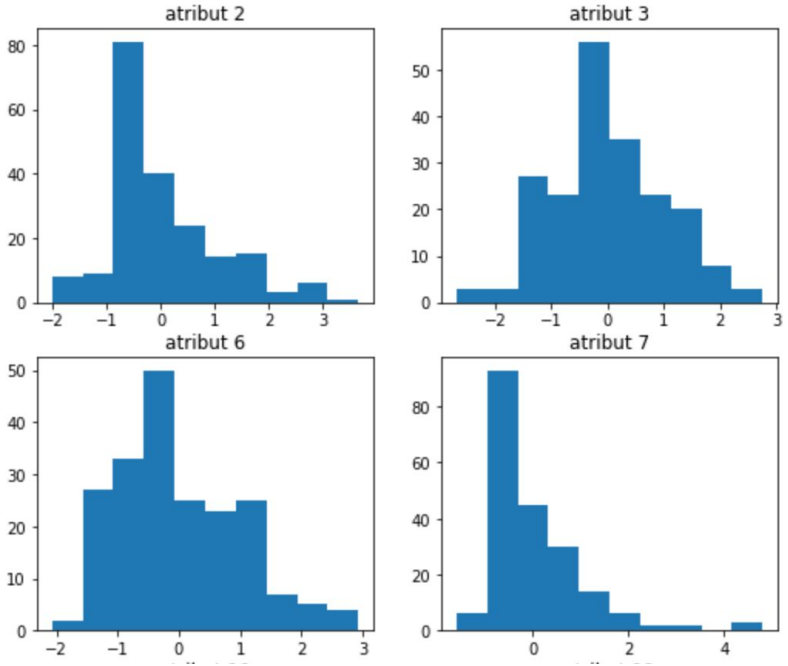
El mejor atributo es “engine_size”, el cual tiene un MSE inferior al resto de atributos.



Histogramas de los atributos

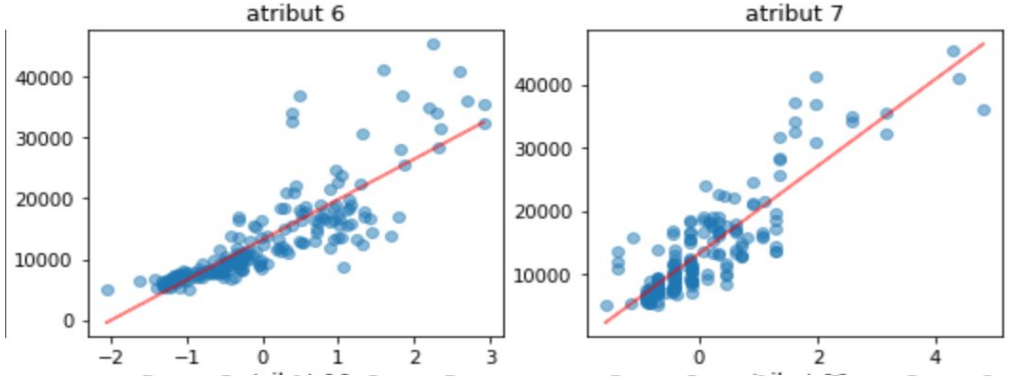
Histogramas con atributos normalizados.

- Nos ayudan a ver las distribuciones gracias a que la media es 0 y la varianza 1.
- En nuestro caso consideramos que todos tienen distribución gaussiana aceptable y no descartamos atributos

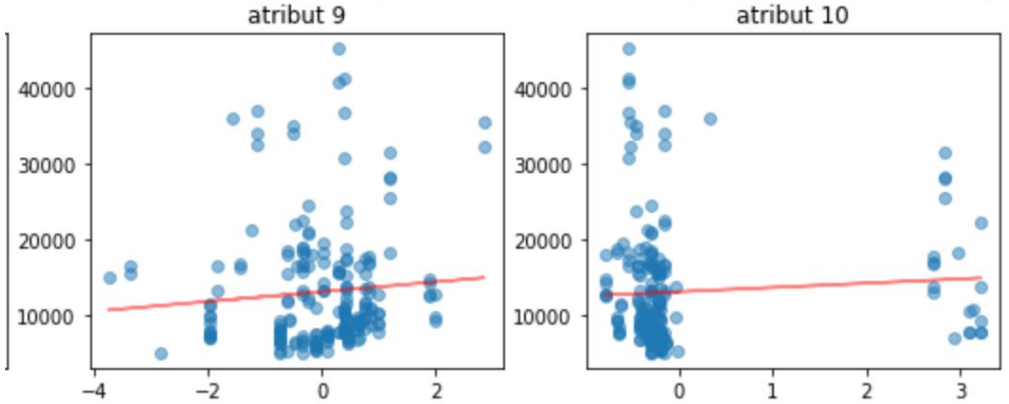


Primeras Regresiones

Atributos que no aportan información:



Atributos que solo aportan ruido:



Resultados

Resultados de la regressión

con todos los atributos

```
media error:  42371307.656414896
media r2:  0.265094211672302
```

Resultados de la regressión

eliminando los atributos ruido

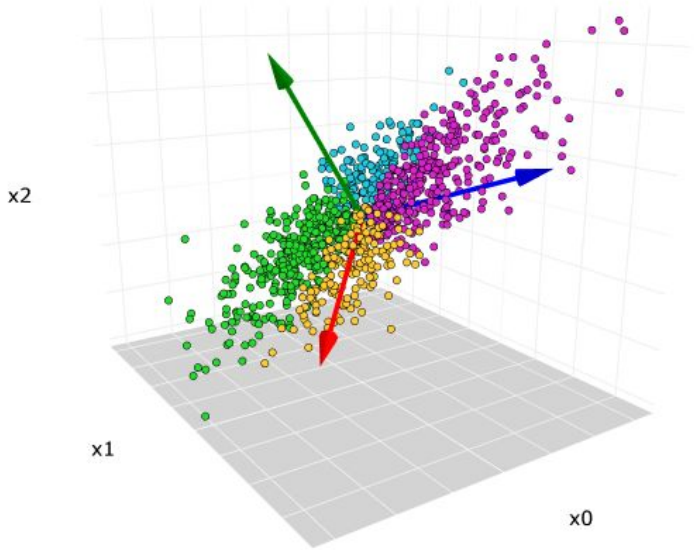
```
media error:  82656812.92281465
media r2:  -0.4336345423717298
```

PCA

Solución para cuando tenemos muchos atributos, reduce la dimensión de los atributos x para que puedan visualizarse en un espacio menor

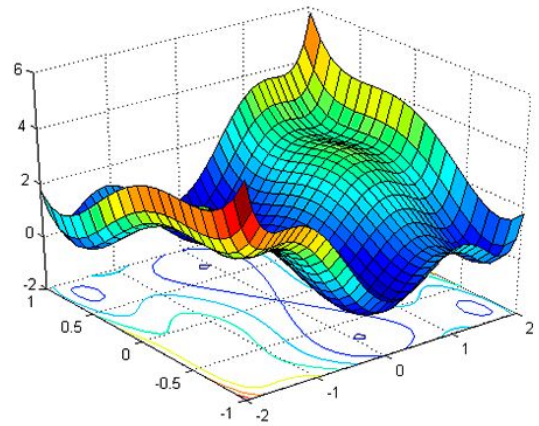
Le asignamos un número de componentes y se ordenan los atributos a partir de la varianza

Objetivo → conseguir la menor pérdida de información a la vez que reducimos las dimensiones de los atributos



Apartado A

Descenso de gradiente



Descenso de gradiente 3D

Fórmula del polinomio

$$f(x) = w_0 + w_1 x^2 + ... + w_{n-1} x^n$$

Fórmula de coste

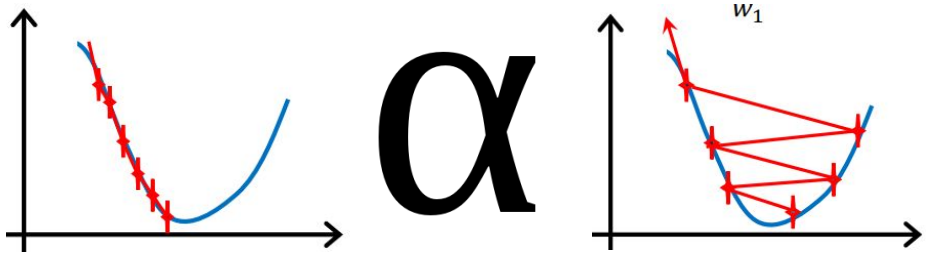
$$J(w) = \frac{1}{2m}\left[\sum_{i=1}^{m}(f(x^i;w) - y^i)^2 + \lambda \sum_{j=1}^{n}(w_j^2)\right]$$

Fórmula para conseguir las Ws

$$w_j := w_j\left(1 - \alpha\frac{\lambda}{m}\right) - \alpha\frac{\lambda}{m}\sum_{i=1}^{m}(f(x^i;w) - y^i)\cdot x_j^i$$

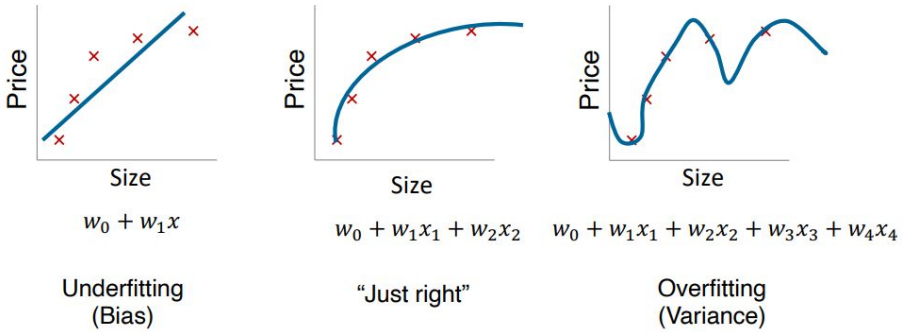
Influencia parámetros

Coeficiente de aprendizaje



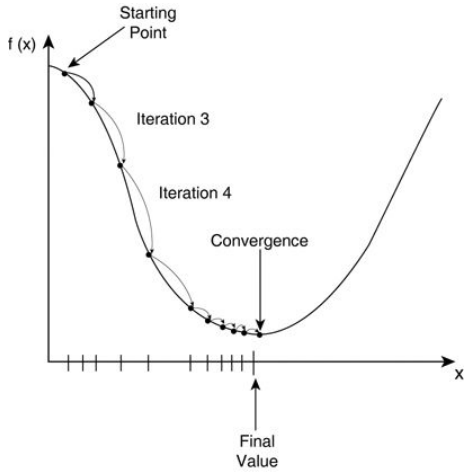
λ

Regularizador



Rapidez de convergencia

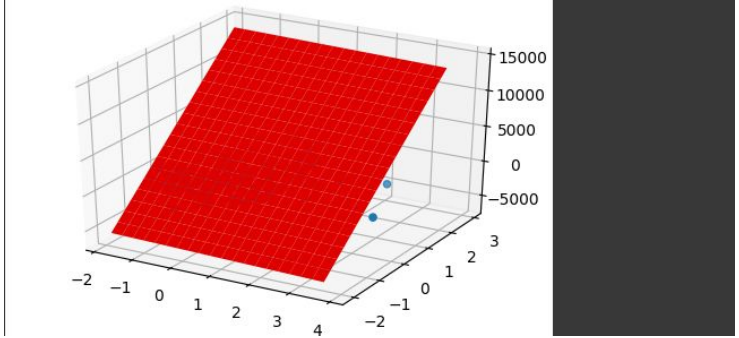
ε



Diferentes funciones polinómicas

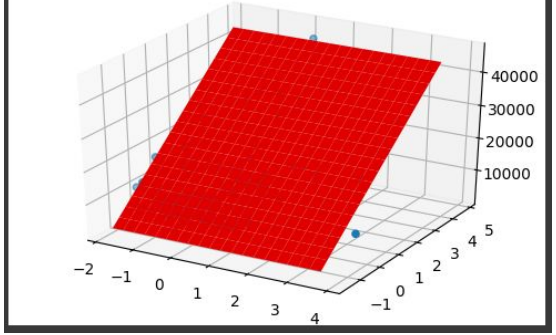
2° Modelo con coste menor

```
Init values: W0: 1.278345158631893 W1: -0.01485764173587552 W2: 0.07538912066139358
Total iterations: 298338
Best cost is: 7_305_803.058891604
Best Ws were: W0: 12244.542439290306 W1: 5444.2462765717555 W2: 901.6687812499653
Attributes selected for W1 and forth were: ['curb_weight' 'engine_size']
<mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7faca3074d50>
```



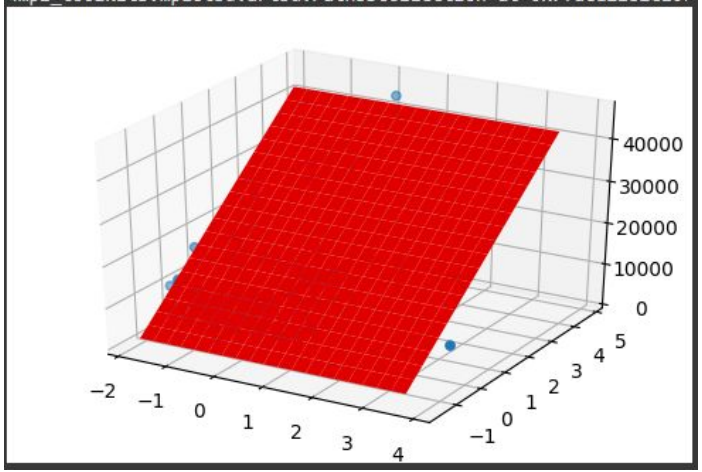
1er Modelo con coste menor

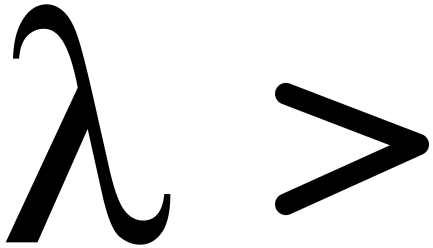
```
Init values: W0: 1.278345158631893 W1: 0.07538912066139358
Total iterations: 206841
Best cost is: 7_518_760.357890498
Best Ws were: W0: 13141.747512438817 W1: 6881.005044669275
Attributes selected for W1 and forth were: ['engine_size']
<mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7faca2b1dc10>
```



Pruebas regularizador

```
Init values: W0: 1.278345158631893 W1: 0.07538912066139358
Total iterations: 32
Best cost is: 12_822_240.686239788
Best Ws were: W0: 12743.008226987606 W1: 6672.207555084623
Attributes selected for W1 and forth were: ['engine_size']
<mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7faca2282c10>
```



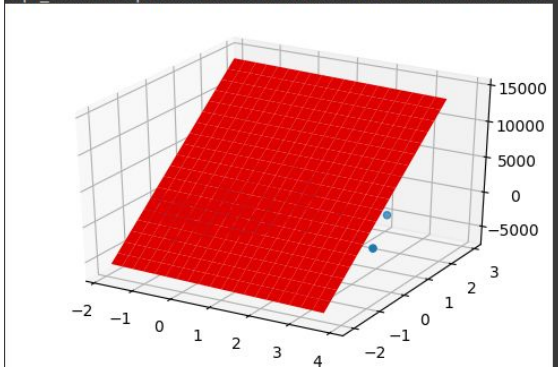


Iteraciones ⇓

Coste ⇑

Regresor manual VS Regresor librería

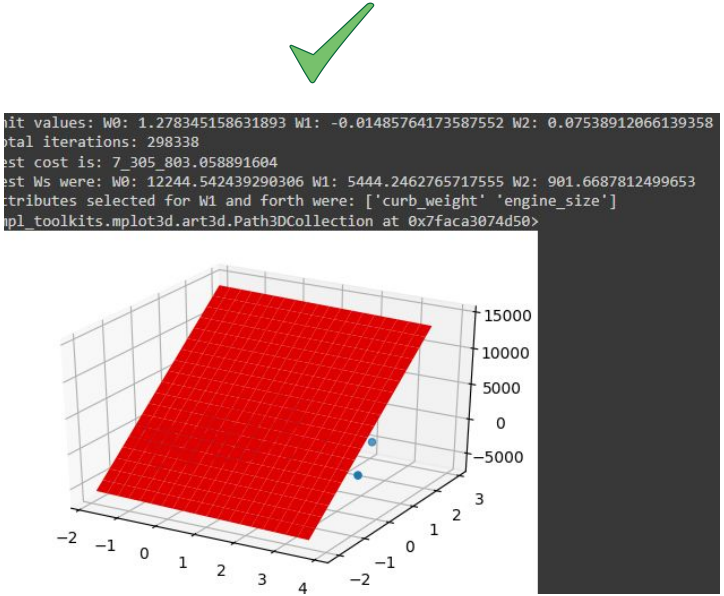
```
Init values: W0: 1.278345158631893 W1: -0.01485764173587552 W2: 0.07538912066139358
Total iterations: 298338
Best cost is: 7_305_803.058891604
Best Ws were: W0: 12244.542439290306 W1: 5444.2462765717555 W2: 901.6687812499653
Attributes selected for W1 and forth were: ['curb_weight' 'engine_size']
<mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7faca3074d50>
```

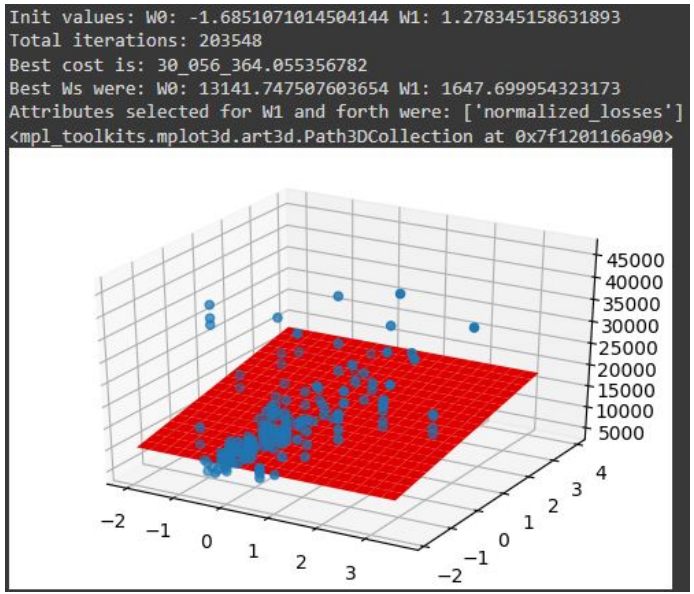



```
MSE atribut  length :  32868224
R^2: atribut  length :  -0.09657864415361583

MSE atribut  width :  27373849
R^2: atribut  width :  0.22820574364110746

MSE atribut  height :  61688099
R^2: atribut  height :  -52.47649857461634

MSE atribut  curb_weight :  19088303
R^2: atribut  curb_weight :  0.5637293416489209

MSE atribut  engine_size :  15021126
R^2: atribut  engine_size :  0.6858854074885816

MSE atribut  bore :  44309278
R^2: atribut  bore :  -1.3909117442321501

MSE atribut  stroke :  62424920
R^2: atribut  stroke :  -148.79544458726366

MSE atribut  compression_ratio :  62523911
R^2: atribut  compression_ratio :  -195.7749562768033
```

Coste ⇊ Más personalizable...

¿Ayuda la visualización?





Muchas Gracias

Ivan Peñarando Martínez
Joel Marco Quiroga Poma
Ferran Martínez Reyes