Introducción a *Machine Learning y* Ciencia de Datos

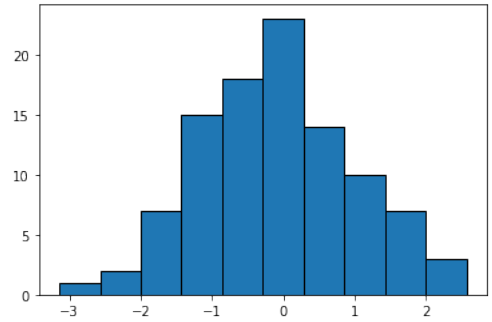
Prueba de normalidad

Para verificar si una variable tiene distribución normal, primero mostremos el resultado con una ficticia creada para este propósito:

```
import numpy as np
data = np.random.normal(0,1,100)
```

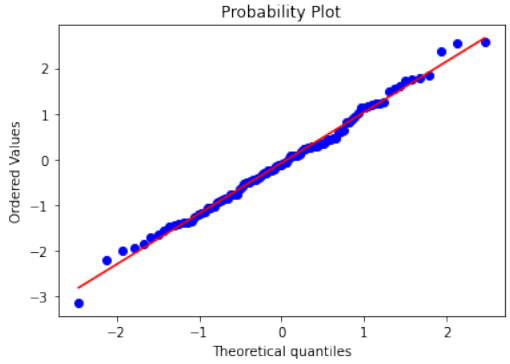
Veamos su histograma:

```
plt.hist(data, edgecolor='black', linewidth=1)
```



Y usemos una gráfica *Quantile-Quantile Normal* con *stats.probplot*:

```
import pylab
import scipy.stats as stats
stats.probplot(data, dist='norm', plot=pylab)
pylab.show()
```



Esta gráfica indica que la distribución es normal si los puntos está distribuídos cerca de la recta. Además, podemos aplicar un contraste de normalidad con Shapiro:

```
from scipy.stats import shapiro
estad, p_value = shapiro(data)
print("Estadístico = %.3f, p_value=%.3f" % (estad, p_value))
# p_value > 0.05 => distribución normal
```

```
Estadístico = 0.993, p_value=0.887
```

Como el p_value es mayor a 0.05, esta prueba indica que la variable tiene distribución normal.

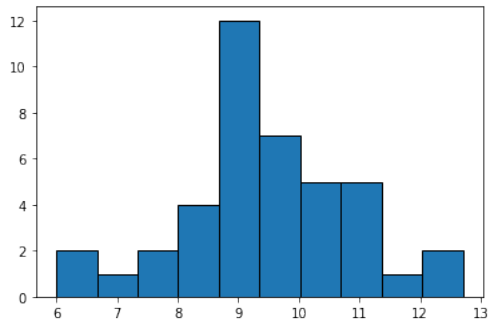
Regresando al conjunto de datos de contaminación, importando bibliotecas útiles:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

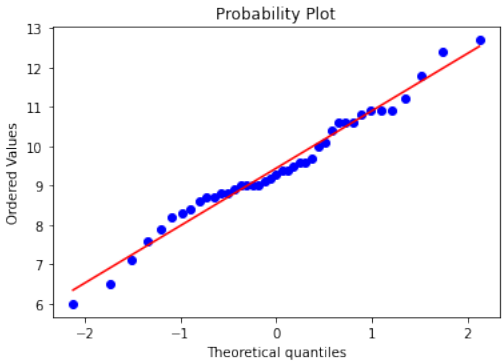
Datos:

```
cont = pd.read_csv('https://bit.ly/31B56KB')
cont.info()
```

```
plt.hist(cont.Velocidad_viento, edgecolor='black', linewidth=1)
```



```
import pylab
import scipy.stats as stats
stats.probplot(cont.Velocidad_viento, dist='norm', plot=pylab)
pylab.show()
```

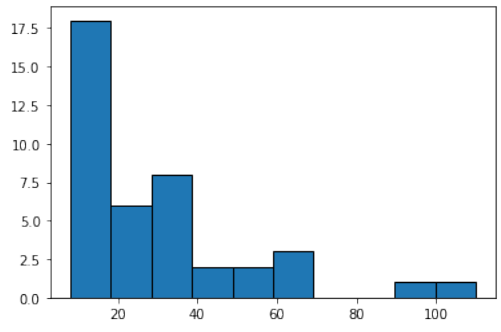



```
from scipy.stats import shapiro
estad, p_value = shapiro(cont.Velocidad_viento)
print("Estadístico = %.3f, p_value=%.3f" % (estad, p_value))
```

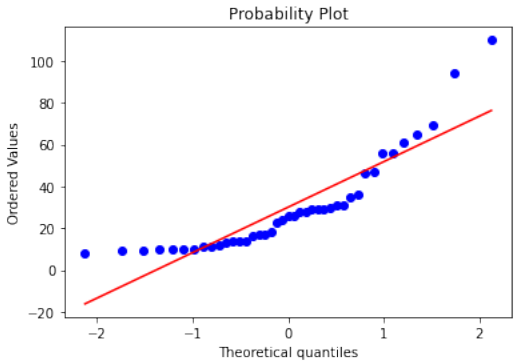
```
Estadístico = 0.981, p_value=0.697
```

Con estos resultados podemos concluir que la variable *Velocidad_viento* tiene distribución normal.

```
plt.hist(cont.Contaminacion_SO2, edgecolor='black', linewidth=1)
```



```
import pylab
import scipy.stats as stats
stats.probplot(cont.Contaminacion_SO2, dist='norm', plot=pylab)
pylab.show()
```



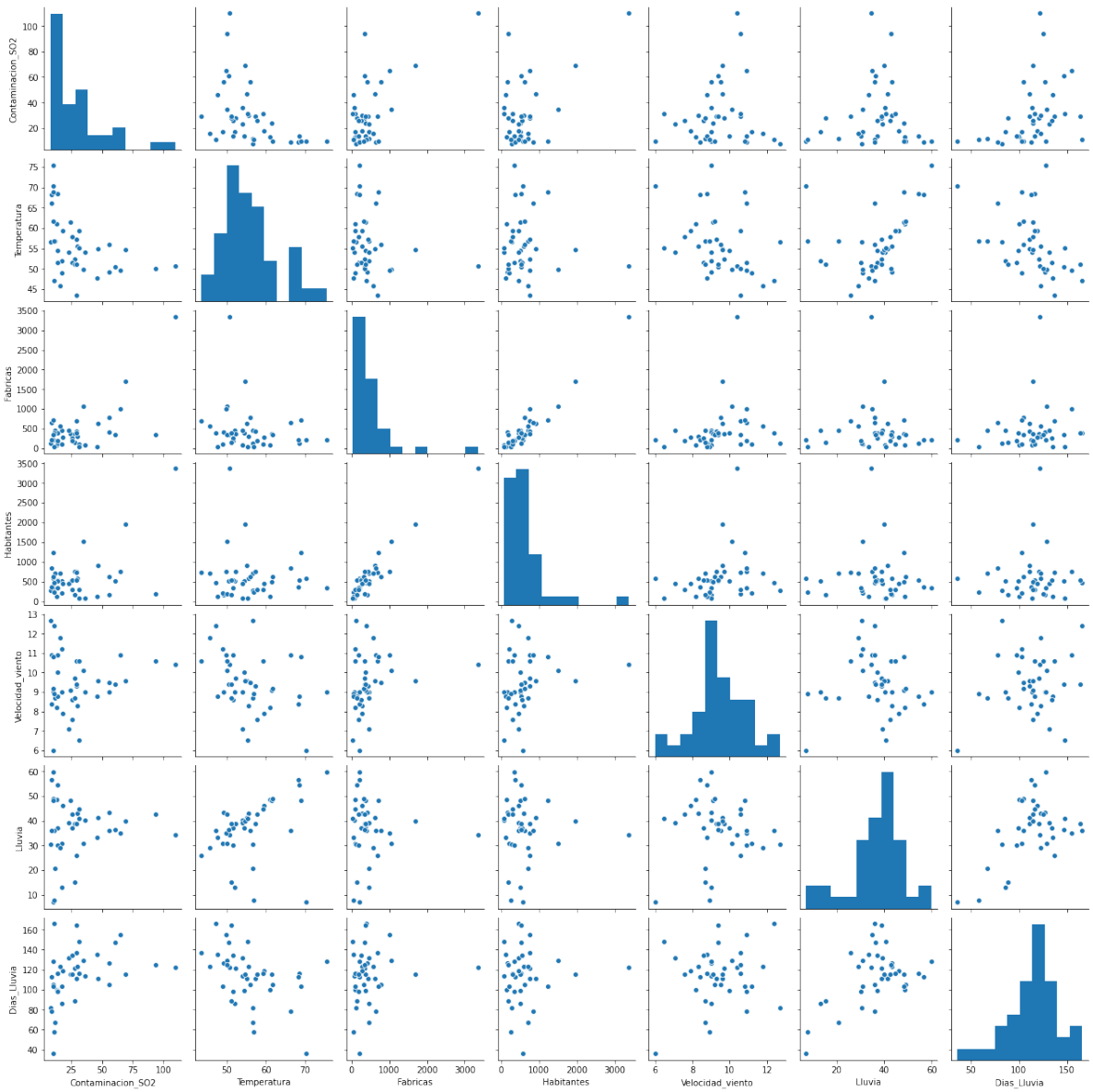

```
from scipy.stats import shapiro
estad, p_value = shapiro(cont.Contaminacion_SO2)
print("Estadístico = %.3f, p_value=%.3f" % (estad, p_value))
```

```
Estadístico = 0.812, p_value=0.000
```

Con estos resultados podemos concluir que la variable *Contaminacion_SO2* **no** tiene distribución normal.

Para ver todas las variables juntos, podemos realizar una diagrama de dispersión por pares de variables:

```
sns.pairplot(cont)
```



En la diagonal, este diagrama muestra el histograma de cada variable, con esto podemos ver si tiene distribución normal.