Limitantes de los coeficientes de correlación lineal

In [ ]:
```
# Importando librerías
import empiricaldist
import janitor
import matplotlib.pyplot as plt
import numpy as np
import palmerpenguins
import pandas as pd
import scipy.stats
import seaborn as sns
import sklearn.metrics
import statsmodels.api as sm
import statsmodels.formula.api as smf
import statsmodels.stats as ss
import session_info
```

Establecer apariencia general de las gráficas

In [ ]:
```
%matplotlib inline
sns.set_style(style='whitegrid')
sns.set_context(context='notebook')
plt.rcParams['figure.figsize'] = (11, 9.4)

penguin_color = {
    'Adelie': '#ff6602ff',
    'Gentoo': '#0f7175ff',
    'Chinstrap': '#c65dc9ff'
}
```

Cargar los datos

Datos Preprocesados

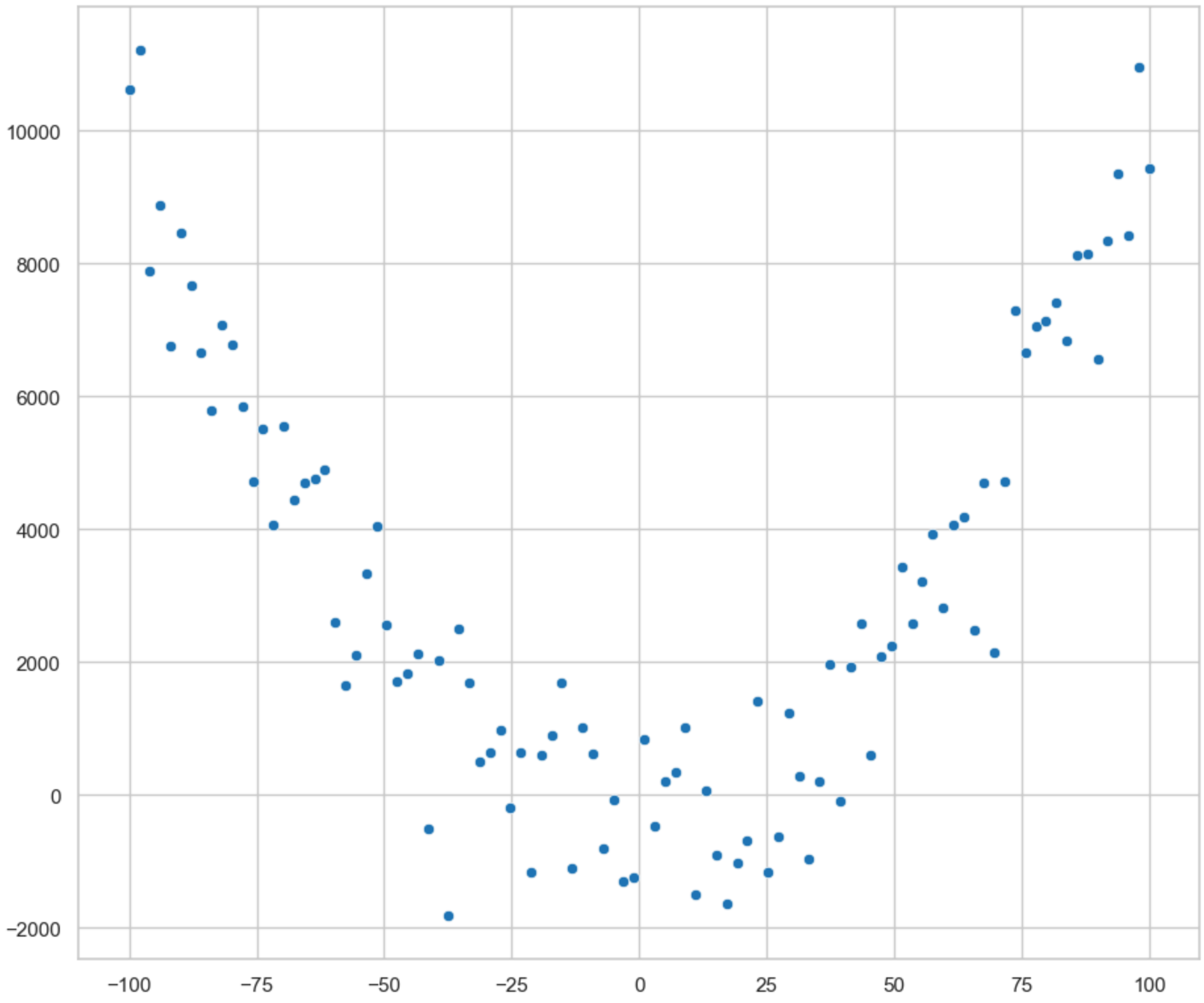
In [ ]:
```
preprocessed_penguins_df = pd.read_csv('dataset/penguins.csv')
```

Limitantes

1.- Solo nos ayuda a determinar la posible existencia de una correlación lineal; sin embargo, su ausencia no significa que no exista otro tipo de correlación

In [ ]:
```
#Valores de x
x = np.linspace(-100,100,100)

#Esta ya no es una correlación lineal
y=x**2

#Añadir ruido para hacerla mas real
y+=np.random.normal(0,1000,x.size)

#Visualizar la gráfica de relación
sns.scatterplot(
    x=x,
    y=y,
)
```

Out[ ]:
```
<AxesSubplot: >
```



Observemos la relación cuadrática que existe.

Si yo aplico un coeficiente de relación. obtendría un valor cercano a cero

In [ ]:
```
#Correlacionando variables independientes
np.corrcoef(x,y)
```

Out[ ]:
```
array([[1.        , 0.00777304],
       [0.00777304, 1.        ]])
```

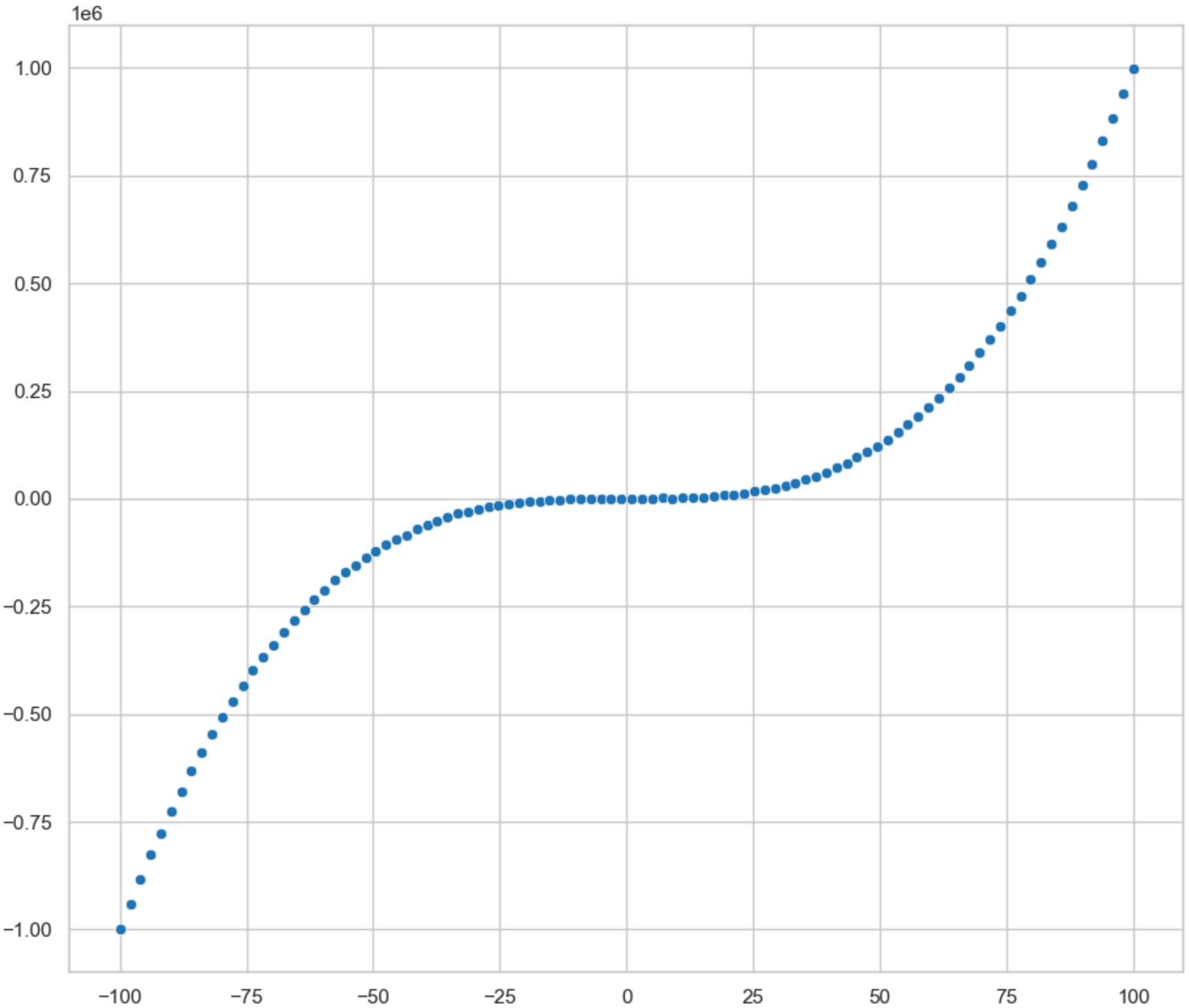
Recordemos que es una matriz con la diagonal en 1, y los demás valores son la correlación entre las demás variables.

De acuerdo al valor de 0.0077, nuestros datos no tienen una correlación lineal, pero no podemos decir que no existe una correlación entre nuestros datos o una asociación entre ellos. Por lo que sabemos que existe una relación cuadrática entre ellos.

Podríamos trabajar de con otros datos que se comportan de diferente manera:

In [ ]:
```
#Valores de x
x = np.linspace(-100,100,100)

#Esta ya no es una correlación lineal
y=x**3

#Añadir ruido para hacerla mas real
y+=np.random.normal(0,1000,x.size)

#Visualizar la gráfica de relación
sns.scatterplot(
    x=x,
    y=y,
)
```

Out[ ]:
```
<AxesSubplot: >
```



Es decir datos que se comporten de manera cubica.

Si yo ejecutará un coeficiente de relación ¿que obtendría?

In [ ]:
```
#Correlacionando variables independientes
np.corrcoef(x,y)
```

Out[ ]:
```
array([[1.        , 0.91674363],
       [0.91674363, 1.        ]])
```

Su valor es cercano a 1, y tienden a ser 1, es decir como le aplicamos el coeficiente de correlación nos podría decir que está relacionado positivamente, pero nosotros sabemos que los datos los tomamos de una función cubica.

Con esto queremos dar a entender que hay veces que tenemos coeficientes altos pero no necesariamente tienen una interacción lineal entre nuestros datos.

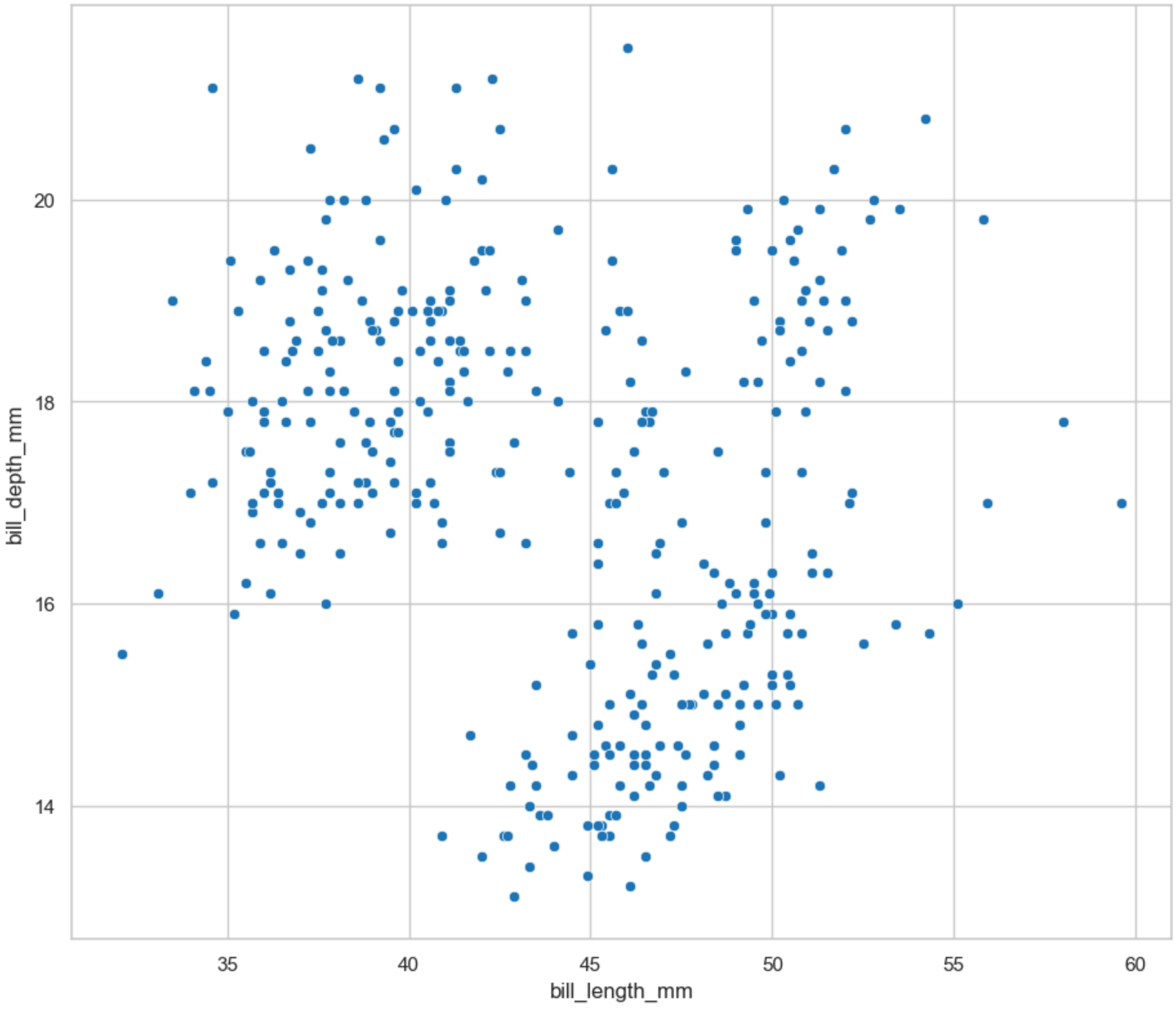
Así que siempre hay que visualizar los datos, mas allá de que tengamos unos coeficientes de correlación que se vean bien.

NOTA; OBSERVAR SIEMPRE NOS VA AYUDAR A ENTENDER A LOS DATOS Y QUE PODRÍA ESTAR CAUSANDO UN COMPORTAMIENTO, ASÍ COMO LA MANERA DE TRABAJARLOS.

¿Existe una correlación entre los siguientes datos?

In [ ]:
```
sns.scatterplot(
    data=preprocessed_penguins_df,
    x='bill_length_mm',
    y='bill_depth_mm'
)
```

Out[ ]:
```
<AxesSubplot: xlabel='bill_length_mm', ylabel='bill_depth_mm'>
```



Yo creo que no existe una correlación por que los datos están muy dispersos y no se ve una tendencia de comportamiento.

Además podemos apoyarnos del coeficiente de correlación.

In [ ]:
```
preprocessed_penguins_df.corr()
```

```
/tmp/ipykernel_251704/3794149396.py:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a futur
e version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  preprocessed_penguins_df.corr()
```

Out[ ]:

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	year
bill_length_mm	1.000000	-0.235053	0.656181	0.595110	0.054545
bill_depth_mm	-0.235053	1.000000	-0.583851	-0.471916	-0.060354
flipper_length_mm	0.656181	-0.583851	1.000000	0.871202	0.169675
body_mass_g	0.595110	-0.471916	0.871202	1.000000	0.042209
year	0.054545	-0.060354	0.169675	0.042209	1.000000

In [ ]:
```
for column in preprocessed_penguins_df.columns:
    print(column)
```

```
species
island
bill_length_mm
bill_depth_mm
flipper_length_mm
body_mass_g
sex
year
```

2.- El coeficiente de correlación no nos habla del impacto de la relación

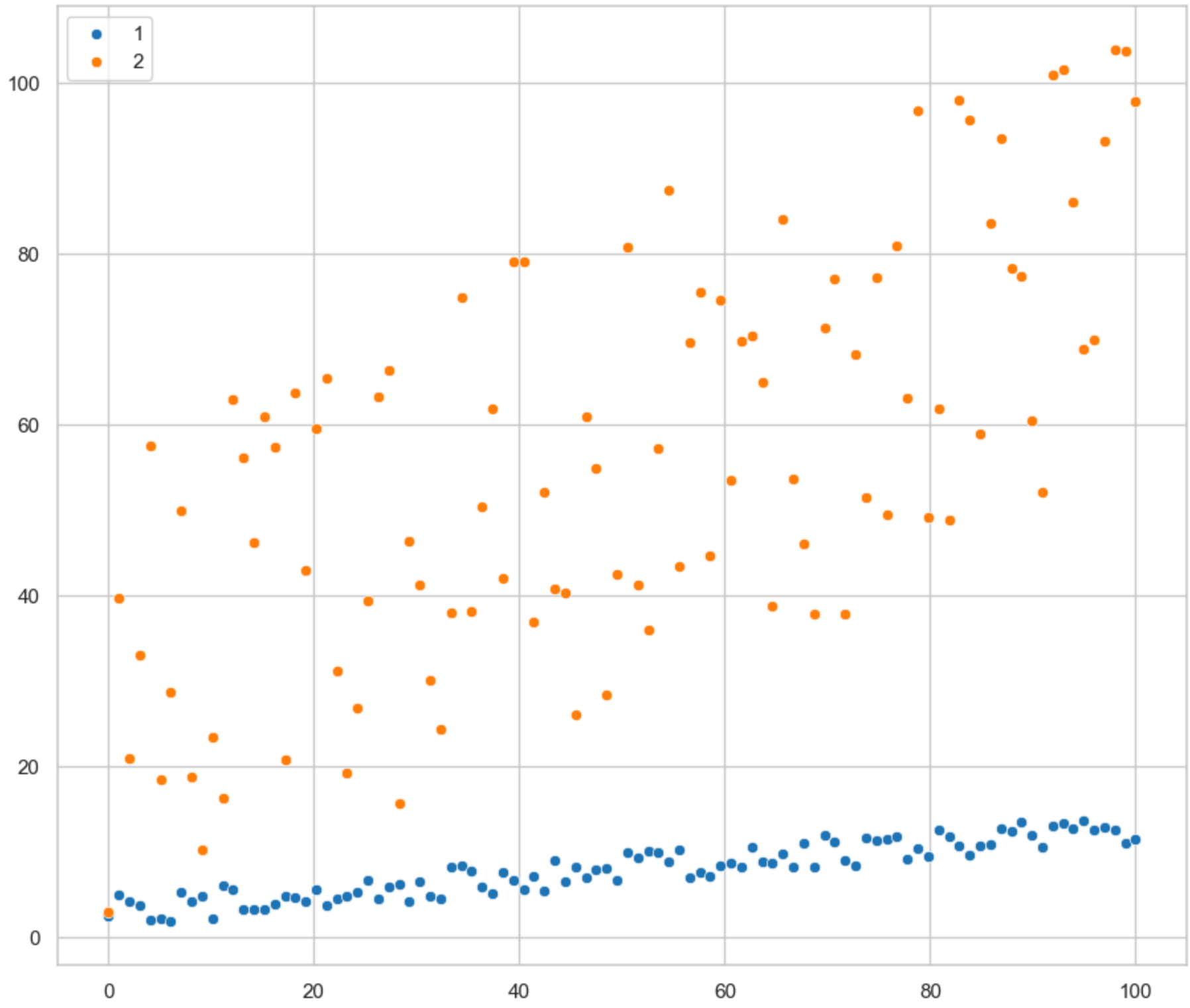
Con impacto nos referimos a que, un coeficiente alto no significa que sea mejor o sirva mejor para predecir algo que podamos implementar para mejor en nuestra empresa.

Ejemplo

In [ ]:
```
np.random.seed(42)
x_1 = np.linspace(0, 100, 100)
y_1 = 0.1 * x_1 + 3 + np.random.uniform(-2, 2, size=x_1.size)
```

```
sns.scatterplot(
    x=x_1,
    y=y_1
)

x_2 = np.linspace(0, 100, 100)
y_2 = 0.5 * x_2 + 1 + np.random.uniform(0, 60, size=x_2.size)

sns.scatterplot(
    x=x_2,
    y=y_2
)

plt.legend(labels=['1', '2'])

print(np.corrcoef(x_1, y_1))
print(np.corrcoef(x_2, y_2))
```

```
[[1.         0.92761617]
 [0.92761617 1.        ]]
[[1.         0.67476343]
 [0.67476343 1.        ]]
```



Nosotros tenemos 2 series:

- azules
- naranjas

Vemos que el coeficiente de la linea azul es de 0.92, es decir un coeficiente de relación casi perfecto y alto.

Después tenemos otro, en el que los puntos están mas dispersos y su coeficiente de relación es de 0.67.

Si nos vamos solo con el coeficiente de relación, nos perdemos de información de utilidad.

Veamos:

Imaginemos que el comportamiento de las series tenga que ver con una variable que controlo y en eje de Y, es el dinero que obtengo al incrementar la variable, entonces:

- es cierto que con la linea azul tenemos un coeficiente de correlación alto, pero el valor de la pendiente que tiene es muy poco o pequeño con referencia a la variable Y.
- Sin embargo la otra serie que está en naranja, la cuantificáramos con respecto del valor del coeficiente de correlación, el resultado es que es muy poco, pero si queremos determinar el impacto que tiene, las X representan un gran cambio; en este caso de dinero.

En casos similares es muy útil ver el impacto y un coeficiente de correlación por si mismo no lo dice.

En la siguiente clase determinaremos el impacto que tienen en la correlación