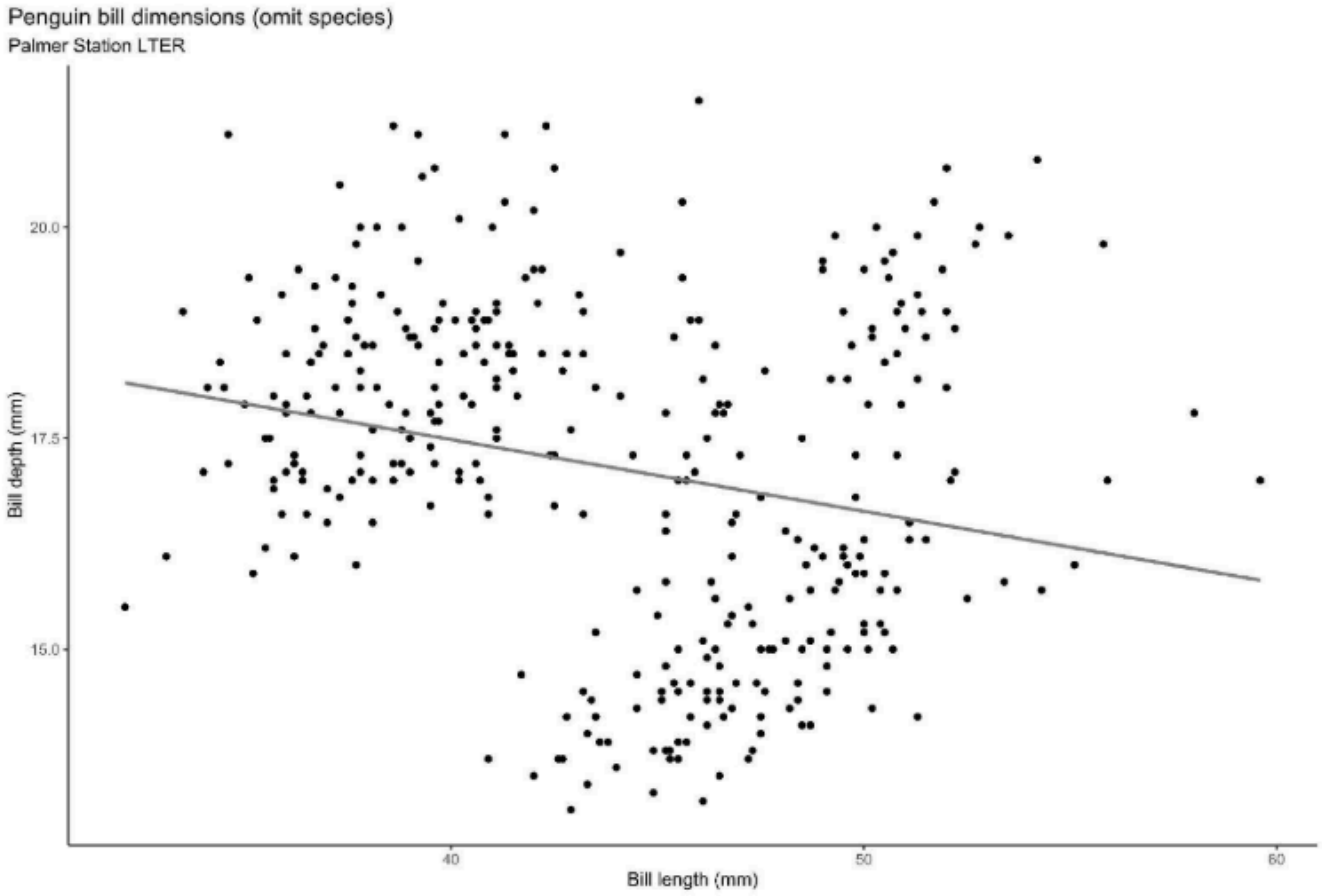
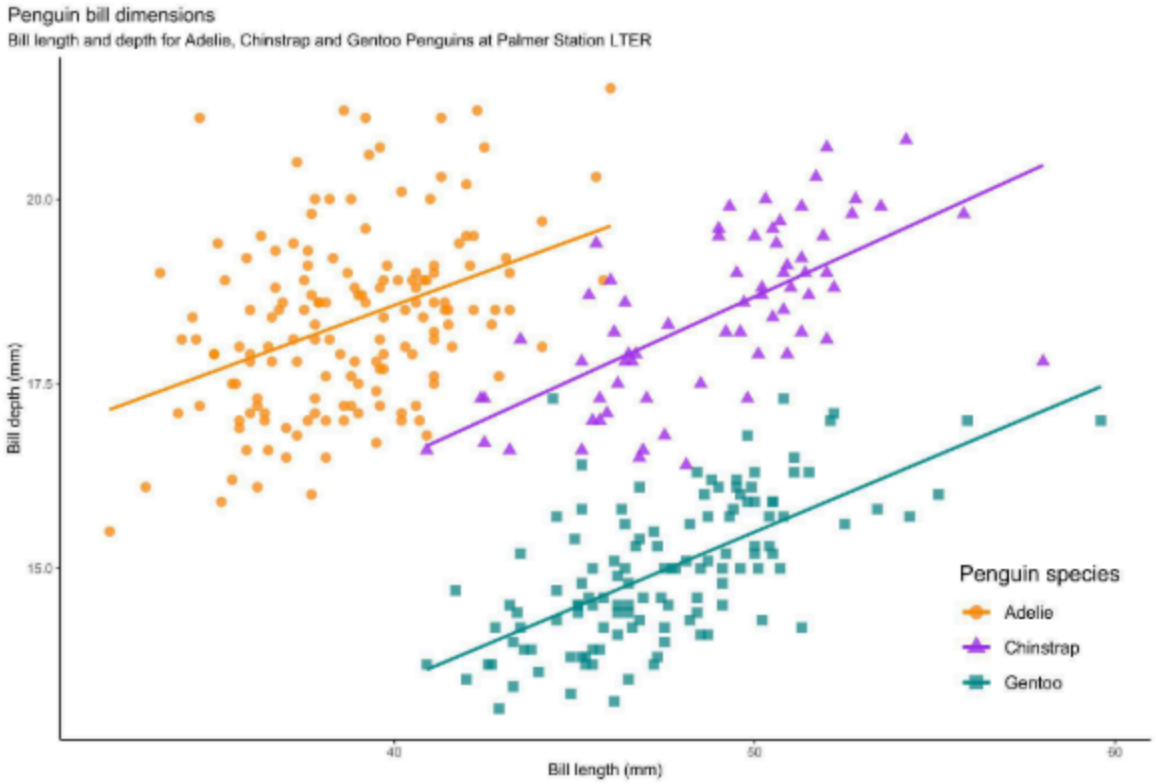
Paradoja de Simpson

A lo largo del curso hemos aprendido a analizar una variable, pares de variables y multiples variables. Así incluirlas poco a poco en el análisis, hacer este proceso de exploración de datos cada vez será fundamental; la razón de esto es que llegaremos a paradojas que nos confundirán y que pueden arruinar todo nuestro trabajo, simplemente porque olvidamos explorar una de nuestras variables.



Por ejemplo nosotros a lo largo del modulo nos enfocamos en las variables de la gráfica. Y nosotros veíamos que tenían una relación negativa es decir; que a mayor longitud de pico nosotros esperábamos tener una altura de pico más pequeña, pero ¿realmente esto nos hacía sentido en la realidad? y si nos hacía sentido o no, los datos parecían que tenían un comportamiento en que se formaban grupos, arriba, abajo; es decir estaban pasando cosas. La razón por la que analizamos de poco en poco fue para llegar al siguiente momento.



En el que incluimos una nueva variable en el análisis, esta fue; `species` y nos dio una nueva perspectiva de lo que estuvimos viendo y haciendo.

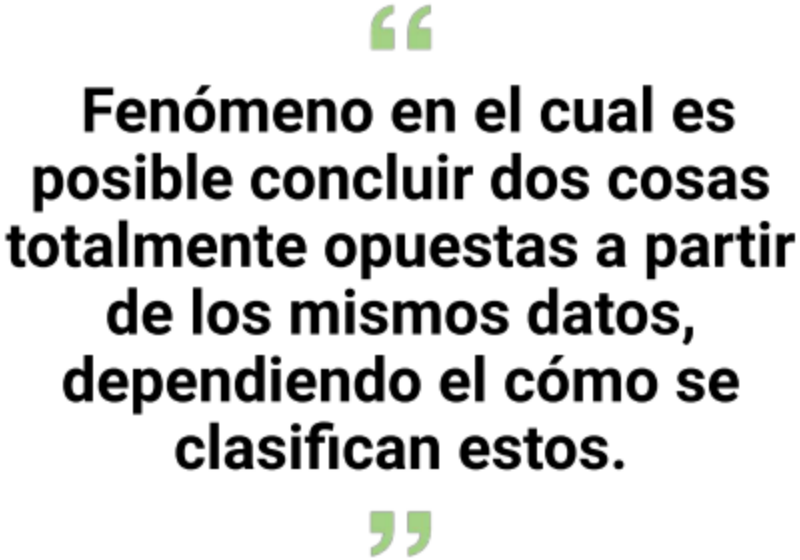
Al agregarla variable `species` nos dimos cuenta que la relación negativa que teníamos, realmente ya no existe, ya no está marcada en nuestros datos. Y ahora tenemos relaciones positivas de los datos, donde es esperado que si la `altura de pico` es mayor, también lo es la `longitud de pico`, donde en términos de proporción todo va a tener sentido.

Ahora para cada especie tenemos diferentes niveles y comportamientos, así que cuando nosotros no incluimos en nuestros modelos este tipo de variables de segmentación, nosotros estábamos omitiendo que las distribuciones podían provenir de diferentes partes.

Recuerda la distribución bimodal que vimos en anteriores clases, cuando explorábamos solo una variable. Ahora intenta explorarla de diferente forma, incluyendo más variables para agrupar o segmentar los datos (gráfica por especies).

Veras como cambian los resultados y todo el panorama al agregar una variable más al análisis, cambia completamente todo lo que habíamos estado analizando. A este fenómeno se le conoce como la paradoja de Simpson.

Paradoja Simpson



Es decir llegar a concluir 2 cosas completamente opuestas a partir de los mismos datos; como cuando nosotros no estábamos clasificando los datos por `especie`, nosotros veíamos que tenían una relación negativa, pero en el momento justo en el que incluimos la variable, la situación cambio completamente.

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

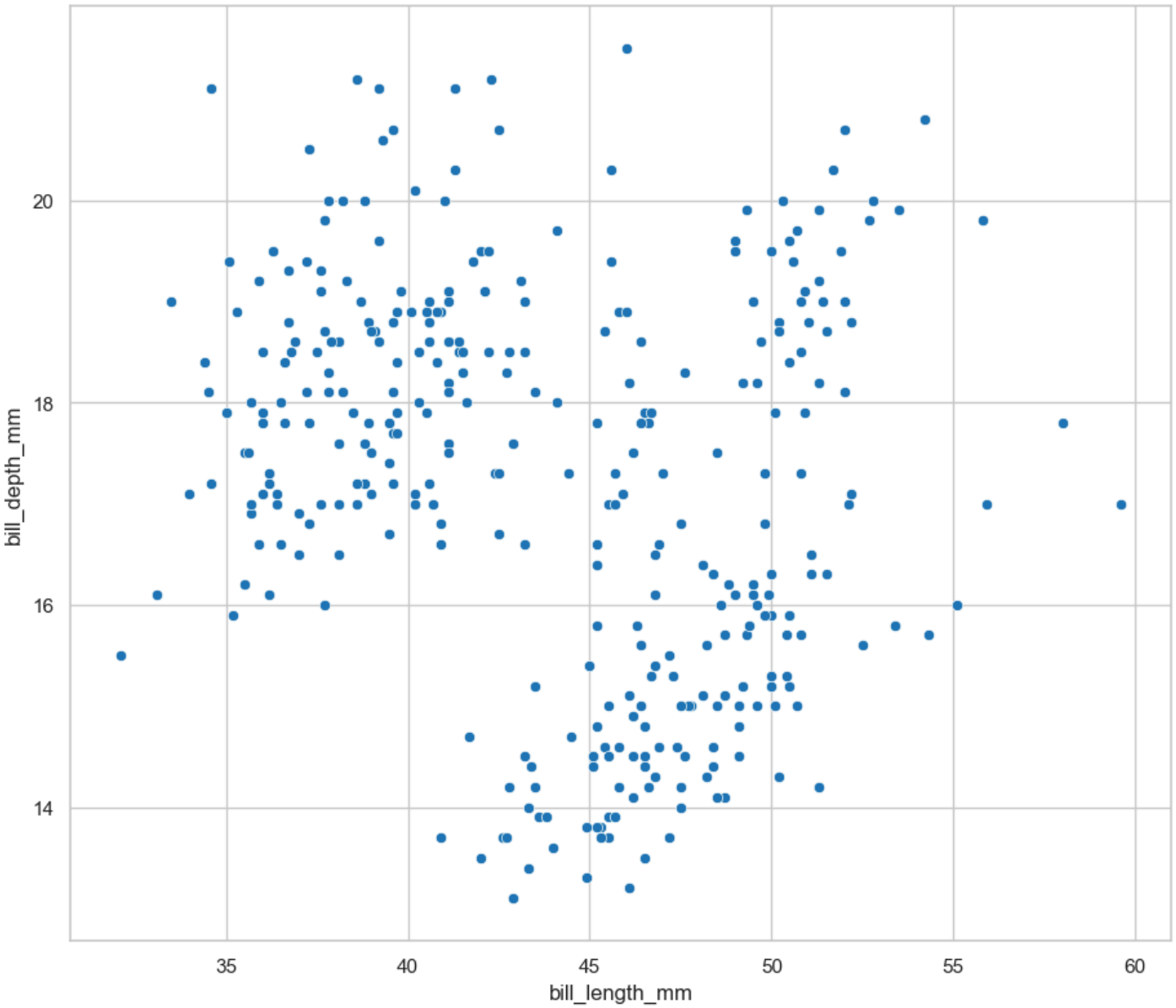
In [ ]:
```
processed_penguins_df = pd.read_csv('dataset/penguins.csv').dropna()
processed_penguins_df.columns
```

Out[ ]:
```
Index(['species', 'island', 'bill_length_mm', 'bill_depth_mm',
       'flipper_length_mm', 'body_mass_g', 'sex', 'year'],
      dtype='object')
```

Primero aprendimos a visualizar los datos en relación uno con otro.

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

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

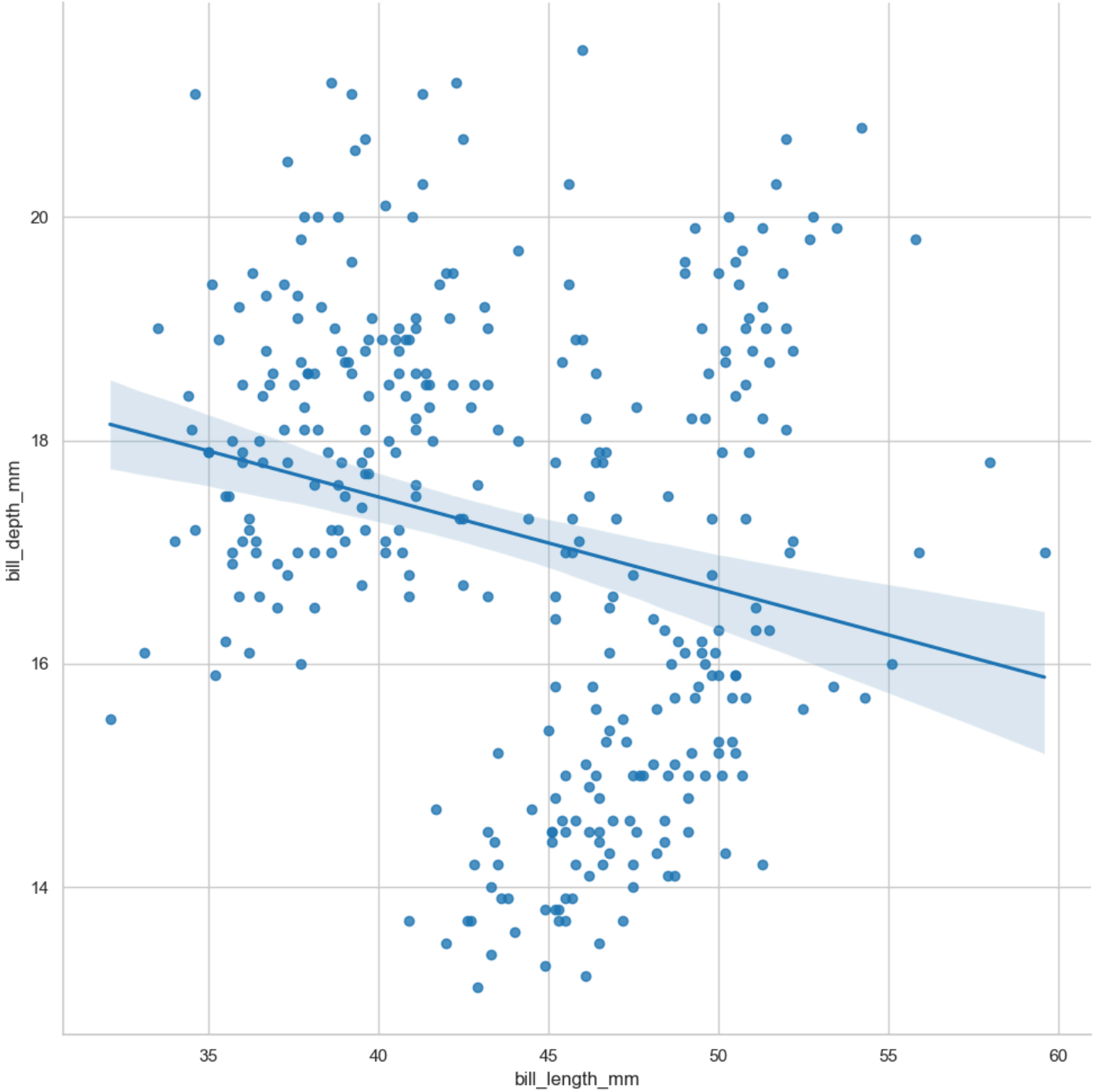


Esto era lo que graficamos y sabíamos en un inicio. Es decir existían comportamientos de separación o huecos, y en esos momentos no contábamos con las herramientas para poder hacer un análisis y saber lo que sucedía.

Después nosotros aprendimos a calcular la regresión lineal simple, para empezar a visualizar la fuerza de la relación de las variables. Llegando a lo siguiente:

In [ ]:
```
sns.lmplot(
    data=processed_penguins_df,
    x='bill_length_mm',
    y='bill_depth_mm',
    height=10
)
```

Out[ ]:
```
<seaborn.axisgrid.FacetGrid at 0x7f14d193af50>
```

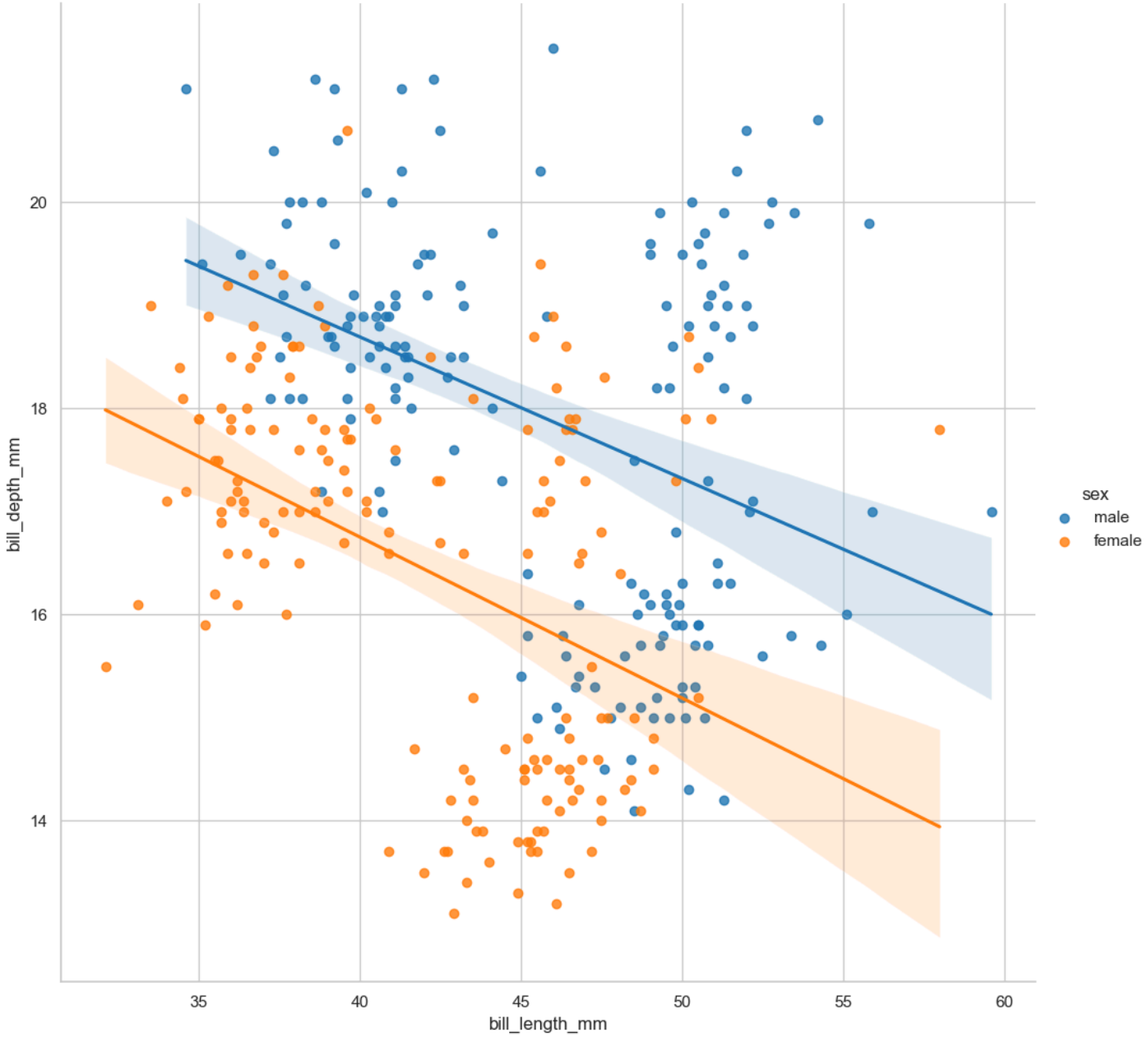


Es decir nosotros observamos una linea y analizamos que conforme; aumentaba la longitud de pico, la profundidad de pico disminuía. Esto realmente no nos convencía o tenia mucho sentido, porque parecía algo raro.

Después incluimos la variable `sexo` y pudimos ver un comportamiento más claro, u otro tipo de variables categóricas.

In [ ]:
```
sns.lmplot(
    data=processed_penguins_df,
    x='bill_length_mm',
    y='bill_depth_mm',
    hue='sex',
    height=10
)
```

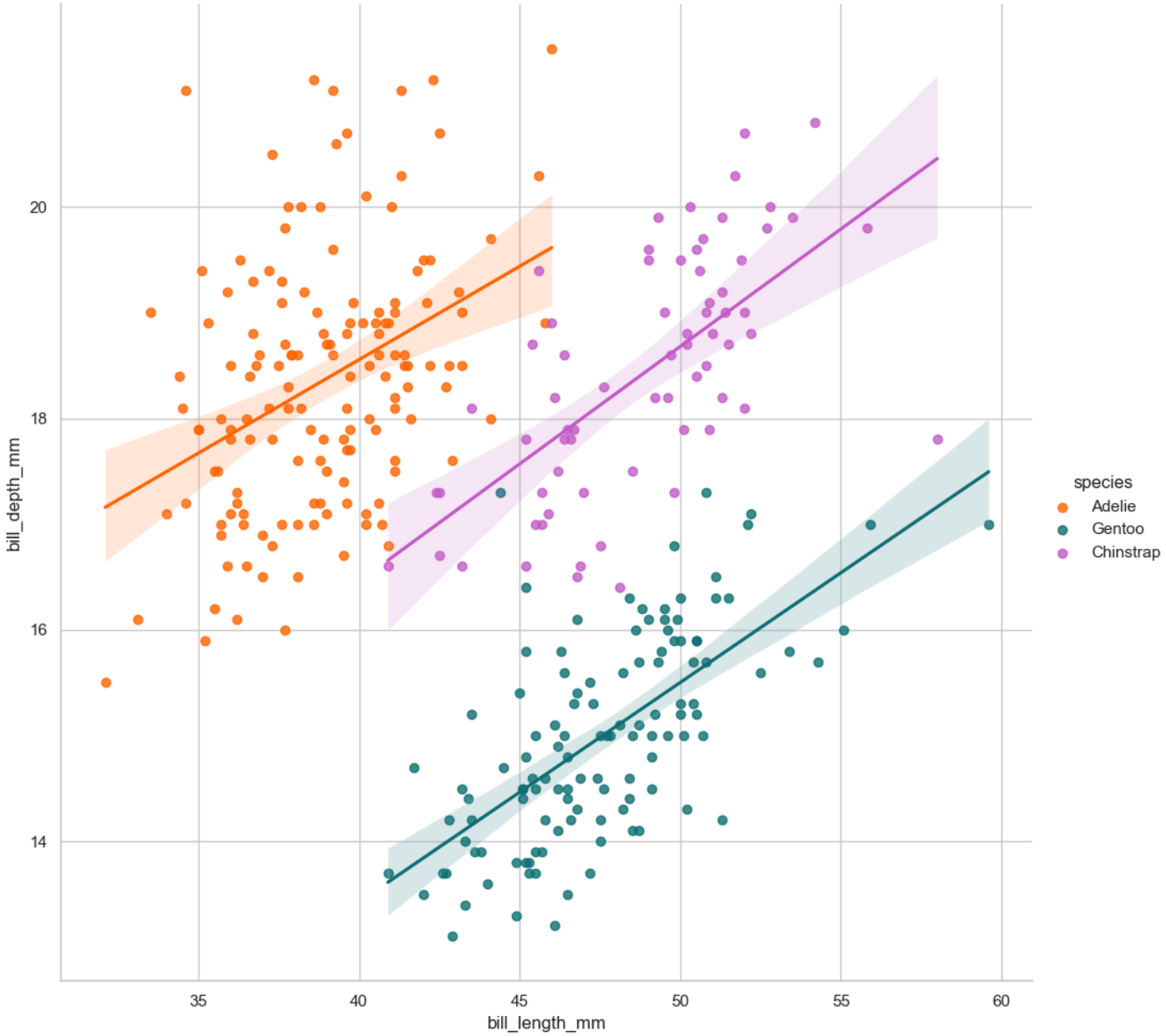
Out[ ]: <seaborn.axisgrid.FacetGrid at 0x7f14d17aa740>



Aunque agregamos la variable `sexo`, no cambio mucho el comportamiento, así que decidimos cambiar la variable categórica.

In [ ]:
```
sns.lmplot(
    data=processed_penguins_df,
    x='bill_length_mm',
    y='bill_depth_mm',
    hue='species',
    palette=penguin_color,
    height=10
)
```

Out[ ]: <seaborn.axisgrid.FacetGrid at 0x7f14d4333070>

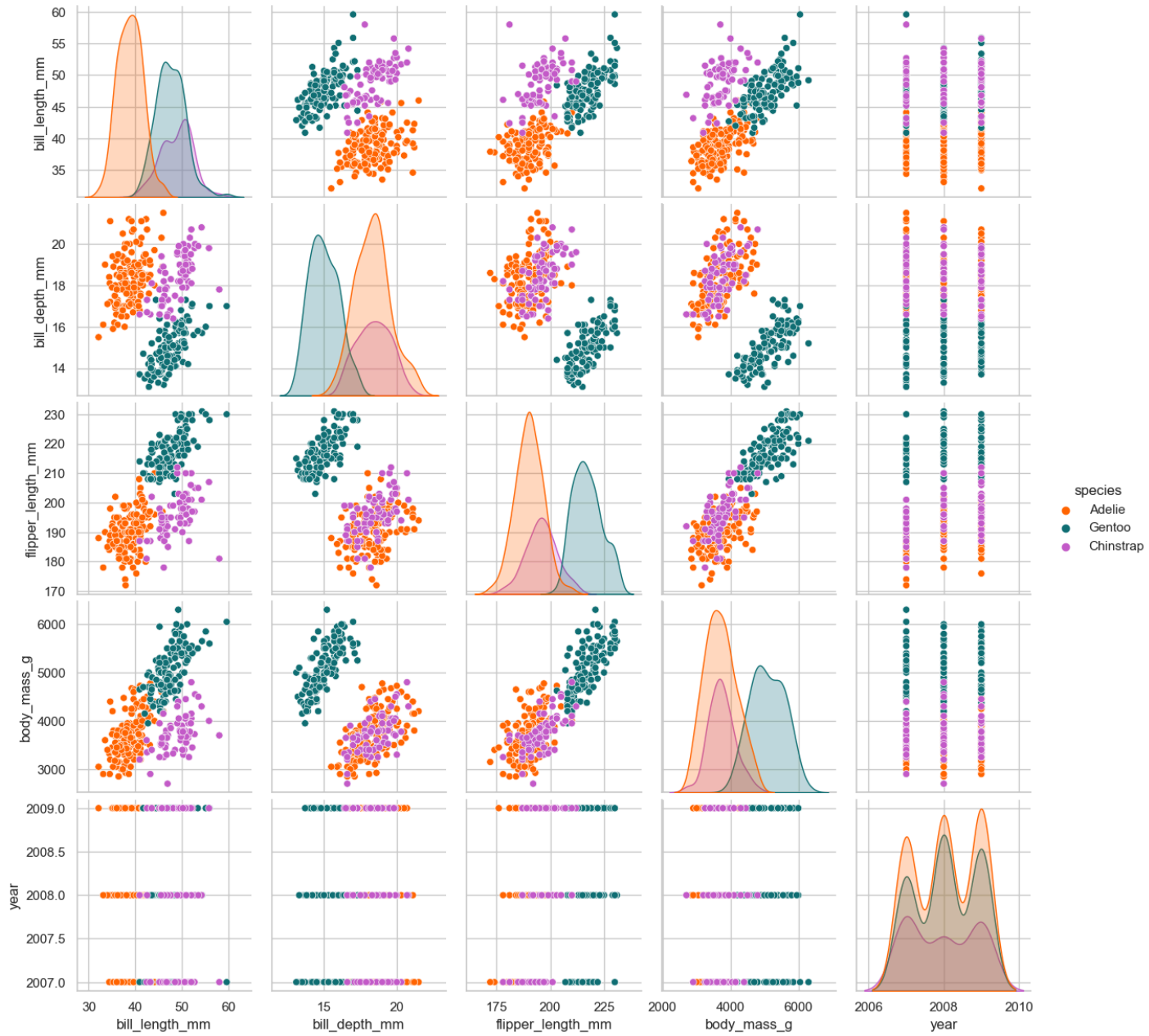


Ahora si, al agregar la variable categórica `species`, logramos visualizar un comportamiento diferente y que nos hace cambiar la forma en que abordamos el problema, ahora en lugar de tener una correlación **negativa**, tenemos una correlación **positiva** para cada una de las especies.

Agregar esta variable nos cambio completamente la forma en que veíamos los datos, por lo que es bueno que volvamos a realizar nuevamente el análisis, considerando las variables categóricas que teníamos en nuestro data set, una forma de hacerlo es a traves de Seaborn por **pairplot**

In [ ]:
```
sns.pairplot(
    data=processed_penguins_df,
    hue='species',
    palette=penguin_color
)
```

Out[ ]:
```
<seaborn.axisgrid.PairGrid at 0x7f14d1746b30>
```



Como podemos observar tenemos las siguientes variables:

- longitud de pico
- profundidad o altura de pico
- longitud de aleta
- masa corporal
- año

Todo esto lo vemos estratificado por especie. En muchas de las graficas vemos que añadiendo una variable, pudimos ver mejor la información. Aunque antes hicimos un modelo basado en una sola variable, no pudimos obtener tanta información. Pero si ahora incluimos una variable más para analizar los datos por `species` se logra apreciar un mejor análisis y podríamos alcanzar a explicar independientemente a cada uno de los pingüinos.

Recuerda que el objetivo es ayudarnos a explorar adecuadamente los datos y hacernos preguntas para conocerlos a detalle y en esta caso el conjunto de datos era de `species` de pingüinos. Y si en ningún

momento hubiéramos considerado la variable `specie`, nos habríamos perdido de una inmensidad de cosas que podría existir dentro de este conjunto de datos.

Así que siempre visualiza los datos, todas las variables por más mínimo que sea utilízala y averigua en que esta afectado o como puede afectar al conjunto de datos y a otras variables

Extras:

- La paradoja de Simpson.