Facultad de Ciencias Económicas
Elements of Machine Learning

Tarea 01
Ajuste de distribuciones y exploración de datos

Daniel Hidalgo
Denisse Bolaños

Objetivo:

Realizar Análisis Exploratorio de Datos y modelar distribuciones en distintos conjuntos de datos con herramientas de visualización y pruebas de hipótesis, para identificar patrones o tendencias en las variables.

Datasets:

Se utilizaron tres datasets distintos: data_aula04, Palmer Penguins y Tips. El primer conjunto de datos contaba con cinco variables, todas numéricas y 677 observaciones; no contenía ningún dato vacío. Palmer Penguins, por otra parte, incluye información de distintas especies de pingüinos encontrados en las “Islas Palmer”. Contiene tres variables categóricas (especies, isla y sexo) y cuatro númericas (largo de aleta, masa corporal, profundidad y largo del pico); con 344 observaciones de las cuales 11 fueron descartadas como tratamiento de datos nulos. Finalmente, “Tips” contiene 244 variables y siete observaciones, cuatro categoricas (sexo, fumadores, día y tiempo) y tres númericas (tamaño de mesa, total de factura y propina); sin necesidad de tratamiento de datos.

Análisis de distribuciones

En este inciso, se analiza el conjunto de datos data_aula04.csv con el objetivo de modelar tres variables seleccionadas mediante distribuciones de probabilidad. Para cada variable, se proponen varias distribuciones, se realizan gráficos de QQ-plot, PP-plot, y se comparan las densidades teóricas y empíricas. Además, se aplica la prueba de hipótesis de Kolmogorov-Smirnov (KS) para determinar cuál de las distribuciones propuestas se ajusta mejor a los datos. El análisis se basa en los resultados de las pruebas KS, que proporcionan una medida cuantitativa del ajuste de cada distribución.

Selección de variables:

var2: Esta variable representa valores numéricos que oscilan entre valores negativos y positivos, con una distribución que parece tener una tendencia central. Se selecciona debido a su variabilidad y potencial para ser modelada por distribuciones simétricas y asimétricas.

var3: Esta variable contiene valores positivos con una dispersión moderada. Se elige por su comportamiento aparentemente sesgado, lo que sugiere que podría ajustarse a distribuciones como la gamma o la log-normal.

var4: Esta variable representa valores numéricos positivos con un rango amplio. Se selecciona por su potencial para ser modelada por distribuciones de cola pesada, como la Weibull o la gamma.

Análisis de distribuciones propuestas:
Cauchy: Se utiliza para modelar datos con colas pesadas, lo que significa que hay una mayor probabilidad de valores extremos en comparación con distribuciones como la normal. Es útil en contextos donde la tendencia central no está bien definida o donde las medias y varianzas no son representativas debido a la dispersión de los datos.

Gamma: Se emplea para modelar datos positivos con sesgo, especialmente cuando los valores pueden ser asimétricos y concentrarse en un rango específico. Es común en el análisis de tiempos de espera, duraciones de eventos y modelado de fenómenos naturales como la radiación o el flujo de partículas.

Weibull (mínima, weibull min): Es una distribución flexible que permite modelar datos con diferentes formas dependiendo de sus parámetros. Se usa frecuentemente en análisis de confiabilidad y tiempos de vida, ya que puede representar desde distribuciones exponenciales hasta distribuciones con colas más pesadas.

Log-Normal: Se aplica cuando los datos son positivos y tienen una distribución sesgada hacia la derecha. Es útil en el análisis de fenómenos como ingresos económicos, tamaños de partículas o precios de activos financieros, donde los valores tienden a multiplicarse en lugar de sumarse.

Pruebas de Kolmogorov-Smirnov:
Se realizaron pruebas de Kolmogorov-Smirnov (KS) para evaluar el ajuste de distintas distribuciones a las variables analizadas. La prueba KS mide la diferencia máxima entre la función de distribución empírica de los datos y la función de distribución teórica. Se selecciona como mejor ajuste la distribución con el menor estadístico KS y un p-value que sugiere una mejor concordancia con los datos observados.

- Para var3, la distribución Cauchy presentó el menor estadístico KS (0.0805) con un p-valor de $2.91x10^{-4}$, lo que indica que, aunque la hipótesis nula es rechazada a niveles de significancia convencionales, es la mejor opción en comparación con las otras distribuciones evaluadas.
- Para var2, nuevamente la distribución Cauchy obtuvo el menor estadístico KS (0.1009) con un p-valor de $1.87x10^{-6}$, lo que la posiciona como la mejor opción para modelar estos datos.
- Para var4, la distribución Weibull Min mostró el mejor ajuste con un estadístico KS de 0. 0.0421 y un p-valor de 0.175, lo que indica que no hay suficiente evidencia para rechazar la hipótesis nula y, por lo tanto, se ajusta bien a los datos.

Conclusiones:
A partir de los resultados obtenidos en las pruebas de Kolmogorov-Smirnov, se pueden extraer las siguientes conclusiones:

1. **La distribución Cauchy fue la mejor opción para modelar var3 y var2:** A pesar de que los p-values son relativamente bajos, la distribución Cauchy mostró el menor error en comparación con Gamma, Weibull Min y Log-Normal, lo que sugiere que captura mejor la naturaleza de los datos en estas variables.

2. **La distribución Weibull Min fue la mejor opción para modelar var4:** A diferencia de las otras variables, var4 mostró un mejor ajuste con la distribución Weibull Min, con un p-valor relativamente alto (0.175), indicando que no hay evidencia significativa para rechazar la hipótesis de que los datos siguen esta distribución.

3. **Las distribuciones Gamma y Log-Normal no ofrecieron buenos ajustes en ningún caso:** En todas las pruebas realizadas, ambas distribuciones presentaron valores de KS más altos y p-values extremadamente bajos, lo que indica que no son apropiadas para modelar los datos analizados.

4. **El análisis destaca la importancia de seleccionar adecuadamente la distribución de probabilidad para cada conjunto de datos:** No todas las distribuciones funcionan bien para todos los casos, y probar varias opciones es importante para encontrar la que mejor representa la realidad. En este caso, la distribución Cauchy fue la mejor para dos de las variables, mientras que la Weibull Min se ajustó mejor a la tercera. Estos resultados pueden ser útiles si en el futuro se quiere hacer predicciones o análisis basados en estos datos.

Palmer Penguins

Conjunto de datos:
El dataset contiene información detallada sobre pingüinos en las Islas Palmer, Antártida. Este incluye mediciones físicas de los pingüinos y datos demográficos de tres especies diferentes: Adelia, Chinstrap y Gentoo. El conjunto consta de 344 registros y 7 variables principales.

Descripción de las Variables:
Las variables incluidas en el conjunto de datos se pueden clasificar en numéricas y categóricas, como se muestra a continuación:

Numéricas:

- bill_length_mm: Longitud del pico del pingüino en milímetros.
- bill_depth_mm: Profundidad del pico del pingüino en milímetros.
- flipper_length_mm: Longitud de las aletas del pingüino en milímetros.
- body_mass_g: Masa corporal del pingüino en gramos.

Categóricas:

- species: Especie del pingüino (Adelie, Chinstrap, Gentoo).
- island: Isla de las Islas Palmer donde fue observado el pingüino (Biscoe, Dream, Torgersen).
- sex: Género del pingüino (Male/Female).

Visualización de datos:
Para el análisis del dataset se usaron técnicas de visualización como histogramas, scatter plots y pairplots. La primera herramienta permitió la exploración y entendimiento de las distribuciones de cada variable. El scatter plot permitió el entendimiento de las relaciones entre cuatro variables: especie, sexo, largo del pico y largo de la aleta. Esta técnica mostró de manera gráfica una posible clusterización de especies basado en estos dos rasgos físicos. Finalmente, el uso de pair plots nos permitió una visualización integral de todas las relaciones entre variables. La aplicación de estas técnicas fueron esenciales para la comprensión y análisis de patrones o tendencias en la data.

Estadística descriptiva:
Con la finalidad de un mayor entendimiento, se realizó un resumen estadístico descriptivo de las variables. Este incluía media, mediana, desviación estándar y cuartiles. El objetivo fue una mayor comprensión de los datos y su comportamiento. Los resultados fueron:

- Masa corporal:
    - Promedio: 4,207.06 g
    - Mediana: 4,050.00 g

- Resumen descriptivo:

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
count	342.000000	342.000000	342.000000	342.000000
mean	43.921930	17.151170	200.915205	4201.754386
std	5.459584	1.974793	14.061714	801.954536
min	32.100000	13.100000	172.000000	2700.000000
25%	39.225000	15.600000	190.000000	3550.000000
50%	44.450000	17.300000	197.000000	4050.000000
75%	48.500000	18.700000	213.000000	4750.000000
max	59.600000	21.500000	231.000000	6300.000000

Conclusiones:
Con el análisis realizado se puede concluir de Palmer Penguins que:

- Existe una variabilidad física según la especie de pingüino. Se puede clasificar por especie según datos como el largo de la aleta y el pico.
- En todas las especies existe una diferencia física según el sexo. Dentro de cada especie, existen diferencias entre hembras y machos a nivel corporal.
- El hábitat podría afectar, aunque en menor medida, en las características físicas de los pingüinos.

Tips

Conjunto de datos:
El conjunto de datos Tips contiene información detallada sobre los pagos en un restaurante, incluyendo el monto total de la cuenta, la propina, características de los clientes (género y si son fumadores o no), el día de la semana, el momento del día en que se realizó la transacción y el tamaño del grupo de comensales. El conjunto de datos cuenta con 244 filas y 7 columnas.

Descripción de las variables
Las variables incluidas en el conjunto de datos pueden dividirse en dos tipos:

Numéricas:

- total_bill: Monto total de la cuenta en dólares.
- tip: Propina dejada en dólares.
- size: Número de personas en la mesa.

Categóricas:

- sex: Género del cliente que pagó la cuenta (Male/Female).
- smoker: Indica si el cliente es fumador o no (Yes/No).
- day: Día de la semana en que se realizó la transacción (Thu, Fri, Sat, Sun).
- time: Momento del día en que se realizó la transacción (Lunch/Dinner).

Visualización de datos:
Para comprender mejor la distribución de los datos y las posibles relaciones entre las variables, se realizaron diferentes tipos de visualizaciones:

Distribución de variables categóricas
Las variables categóricas fueron representadas mediante gráficos de barras, lo que permitió identificar patrones de frecuencia en el conjunto de datos. Se encontraron las siguientes tendencias:

- Género: Se observó una distribución donde hay una mayor frecuencia de clientes masculinos.
- Fumadores: La proporción de fumadores frente a no fumadores, donde los fumadores superan a los no fumadores.
- Día de la semana: La mayoría de las transacciones ocurrieron los fines de semana, especialmente los sábados.
- Momento del día: Hay una mayor cantidad de transacciones durante la cena, lo que podría indicar una tendencia en los hábitos de consumo del restaurante.

Distribución de variables numéricas
Para analizar la distribución de las variables numéricas, se generaron histogramas y funciones de densidad. Los principales descubrimientos fueron los siguientes:

- Total de la cuenta: Se identificó una distribución sesgada a la izquierda, indicando que la mayoría de las cuentas tienen valores relativamente bajos, con algunas excepciones de cuentas significativamente altas.
- Propina: La distribución de propinas también presentó un sesgo similar al del total de la cuenta, lo que sugiere que las propinas suelen estar relacionadas con el tamaño del consumo.
- Tamaño del grupo: Se observó que la mayoría de las mesas estaban compuestas por grupos pequeños de entre 2 y 4 personas.

Además, para explorar diferencias dentro de las variables numéricas, se realizaron comparaciones entre categorías mediante hue en gráficos. Por ejemplo, al analizar la propina separada por género o fumador/no fumador, se pudieron identificar tendencias dentro de cada grupo.

Análisis de relaciones entre variables
Para explorar correlaciones y relaciones entre las variables, se implementaron varios métodos gráficos:

- Scatterplots: Se generaron gráficos de dispersión para examinar la relación entre el total de la cuenta y la propina. Se evidenció una correlación positiva, lo que significa que a medida que el monto de la cuenta aumenta, la propina también suele aumentar.
- Pairplots: Se analizaron las interacciones entre múltiples variables numéricas para detectar patrones más complejos.
- Mapa de calor: Se construyó una matriz de correlación donde se observaron relaciones estadísticas significativas entre ciertas variables. En particular, se encontró una alta correlación entre el total de la cuenta y la propina.

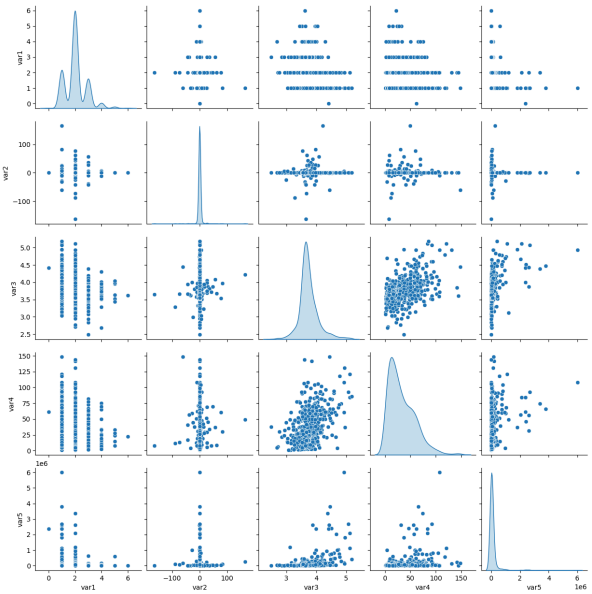
Estadística descriptiva:
Se calcularon estadísticas descriptivas para cada variable, proporcionando una visión general de las tendencias del conjunto de datos. Entre los valores más relevantes se encuentran:

- Propina:
    - Promedio: 3.00 dólares
    - Mediana: 2.90 dólares
- Total de la cuenta:
    - Promedio: 19.79 dólares
    - Mediana: 17.80 dólares

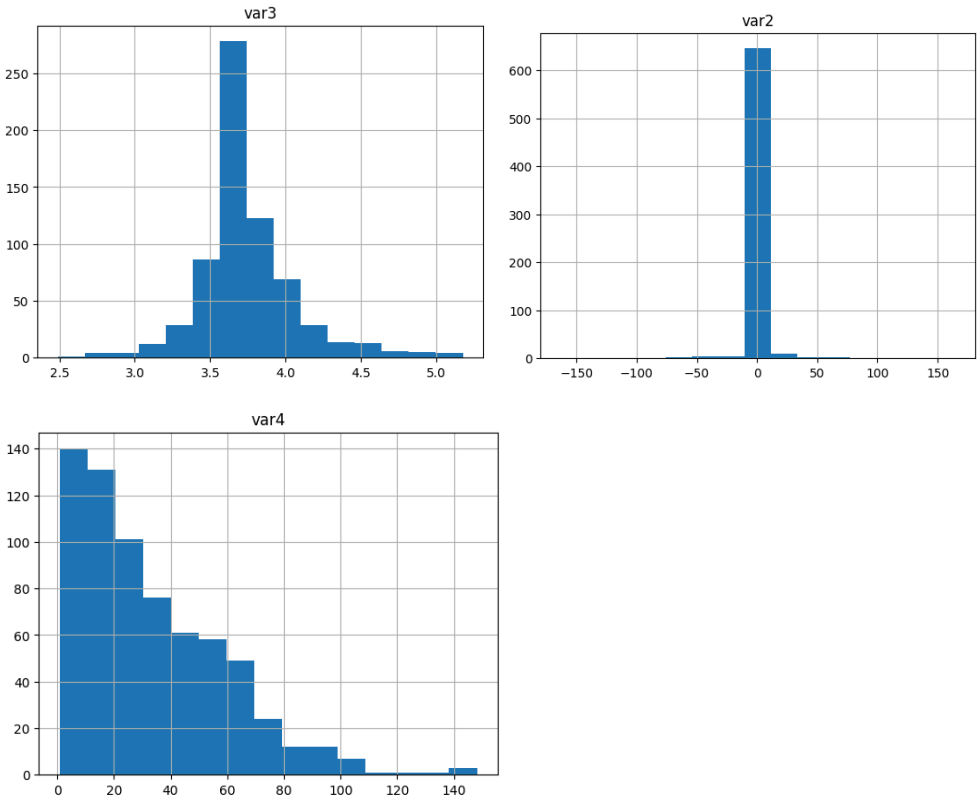
Conclusiones:
A partir del análisis exploratorio del conjunto de datos Tips, se pueden extraer varias conclusiones relevantes:

- Existe una relación positiva entre el total de la cuenta y la propina, lo que sugiere que los clientes suelen calcular sus propinas en función del monto total del consumo.
- Se identificó una mayor actividad en el restaurante durante la cena y los fines de semana, lo que podría estar relacionado con una mayor disposición a gastar y a dejar propinas más elevadas.
- La mayoría de las mesas están conformadas por grupos pequeños, lo que podría influir en la cantidad de propina dejada.

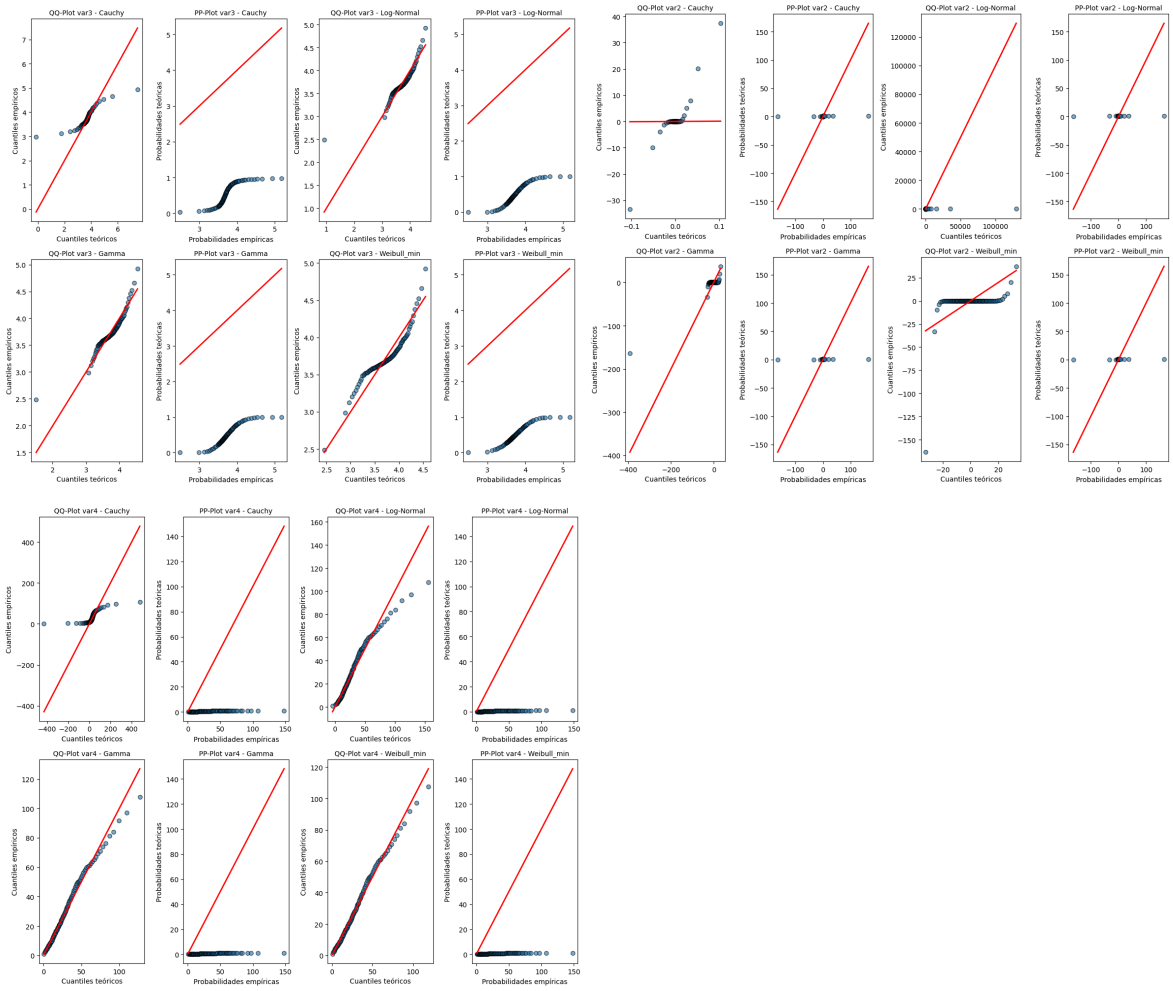
Anexos



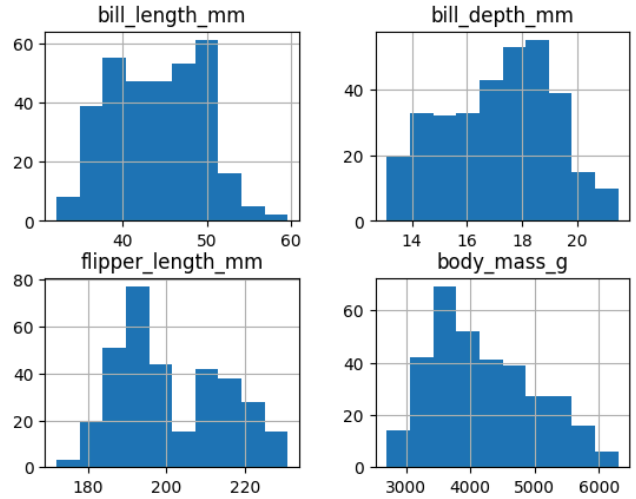
Anexo 1: Pair plot de dataset data_aula04



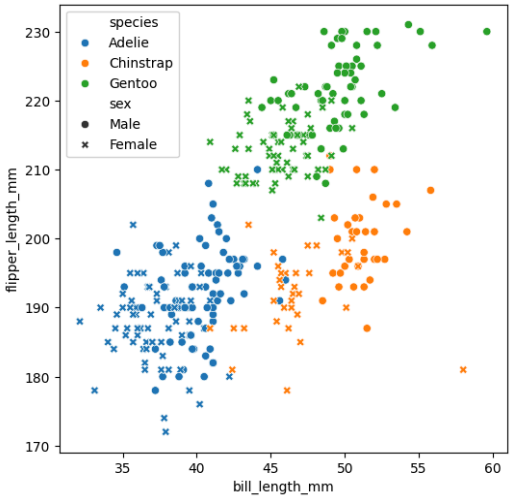
Anexo 2: Histogramas variables data_aula04



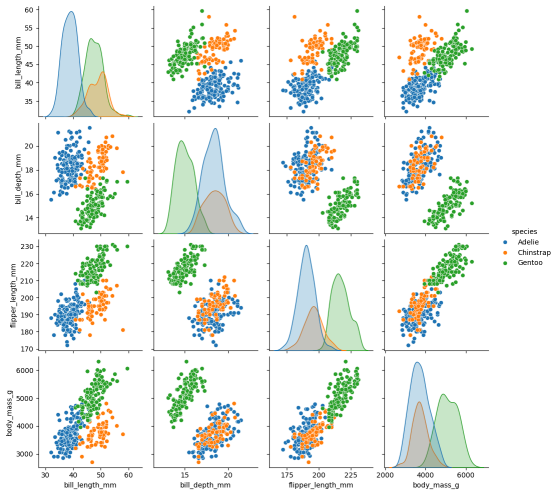
Anexo 3: QQ-Plot y PP-Plot data_aula04



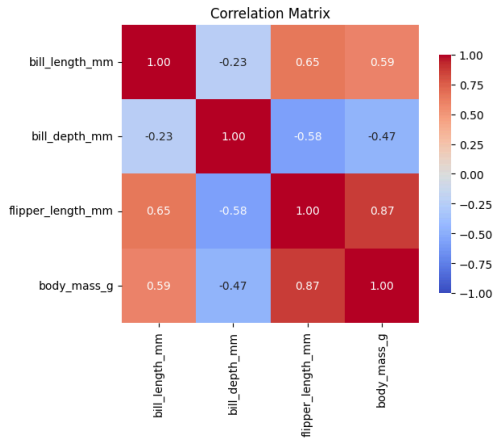
Anexo 4: Histogramas Palmer Penguins



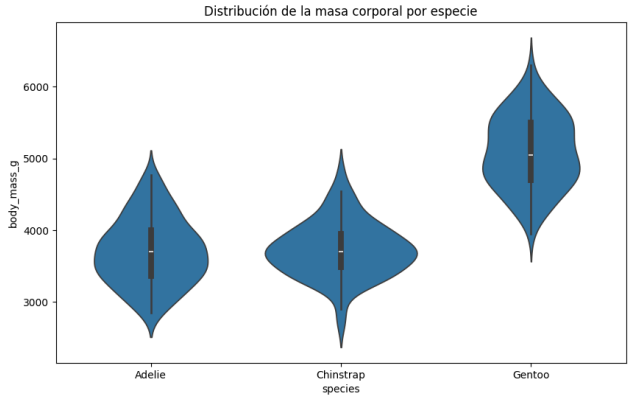
Anexo 5: Scatter Plot de Longitud del Pico vs. Longitud de las Aletas por Especie y Sexo de los Pingüinos



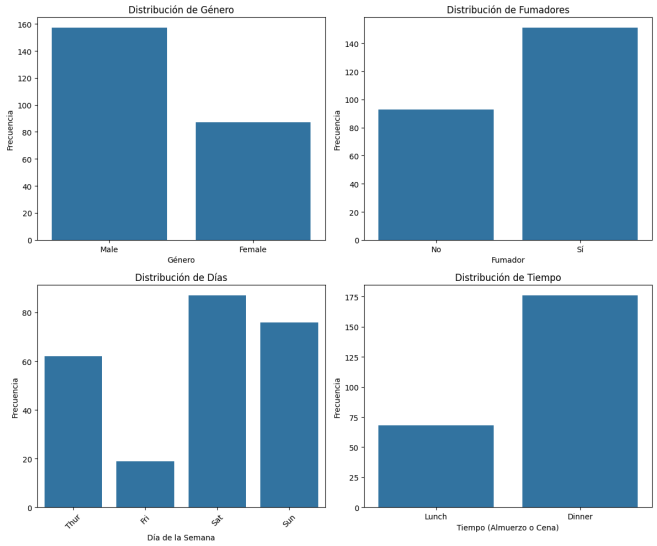
Anexo 6: Pair plot de Distribuciones y Relaciones por Especie para Pingüinos



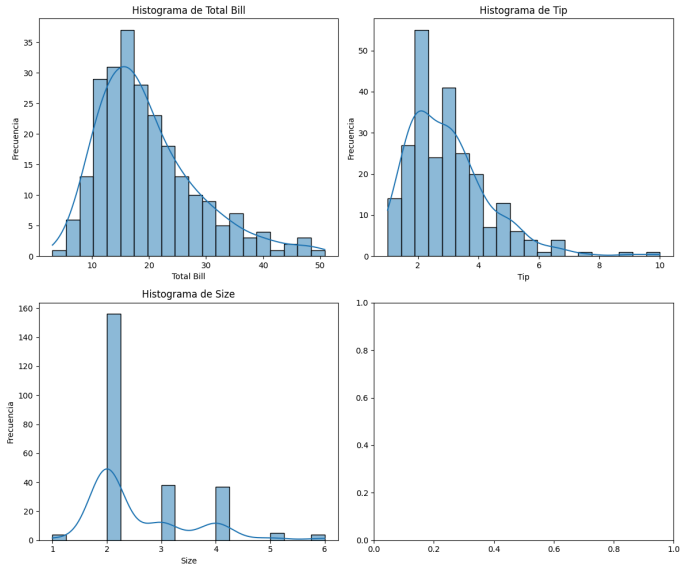
Anexo 7: Matriz de correlación Palmer penguins



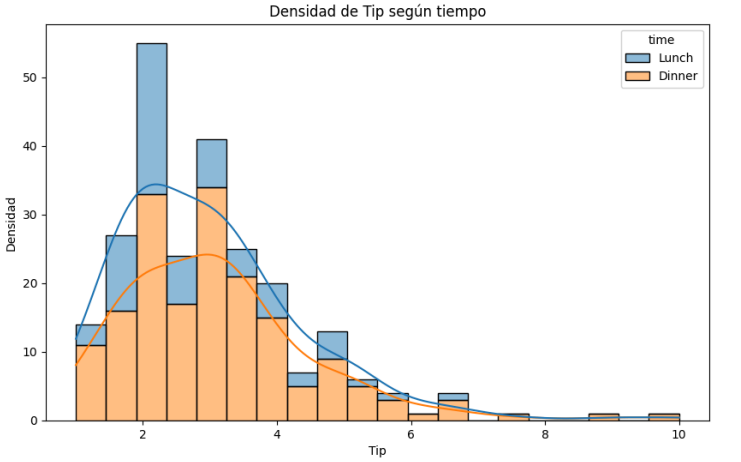
Anexo 8: Distribución de la masa corporal por especie de pingüino

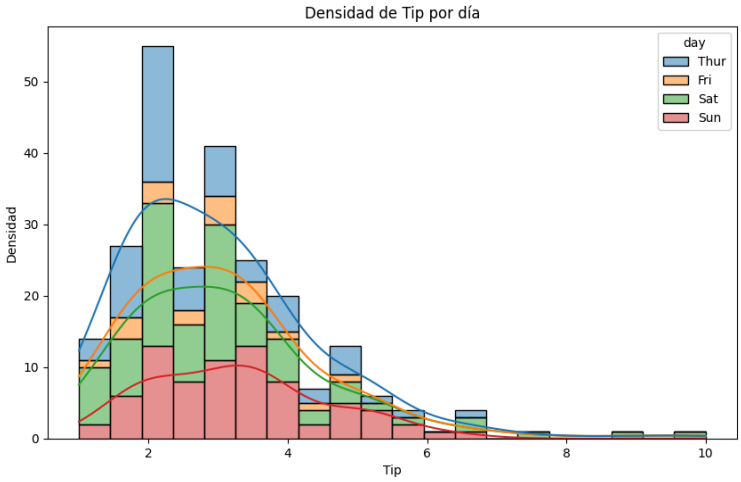


Anexo 9: Distribución por variable categórica Tips

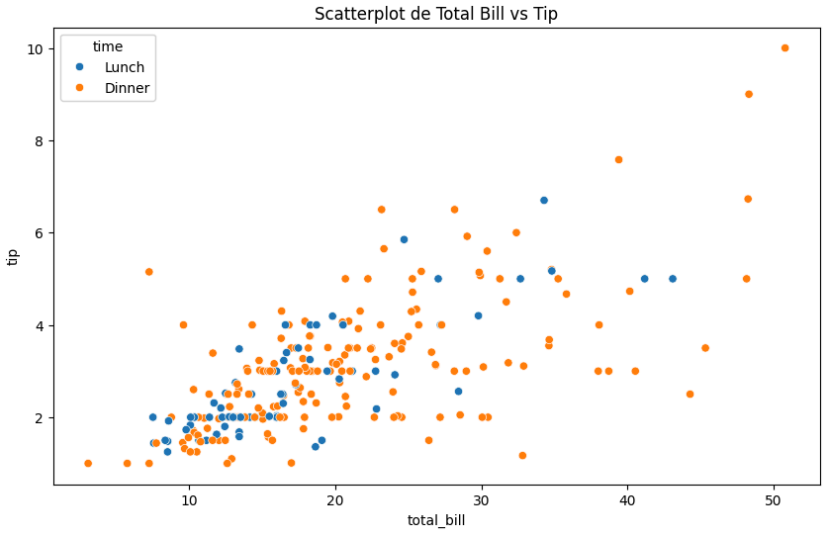


Anexo 10: Histogramas de variables numéricas Tips

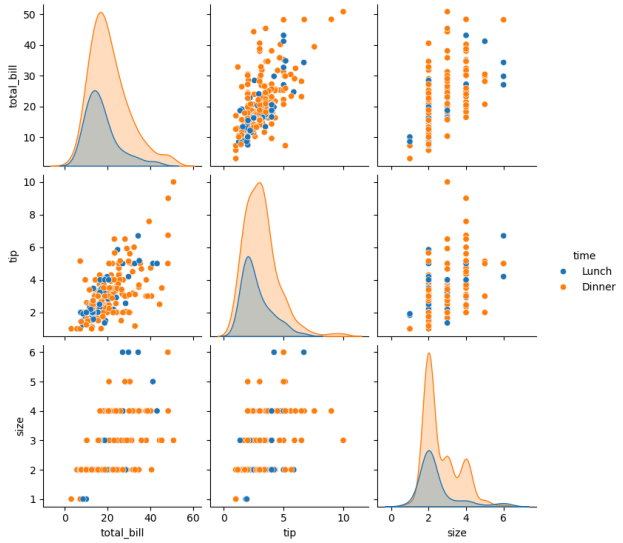




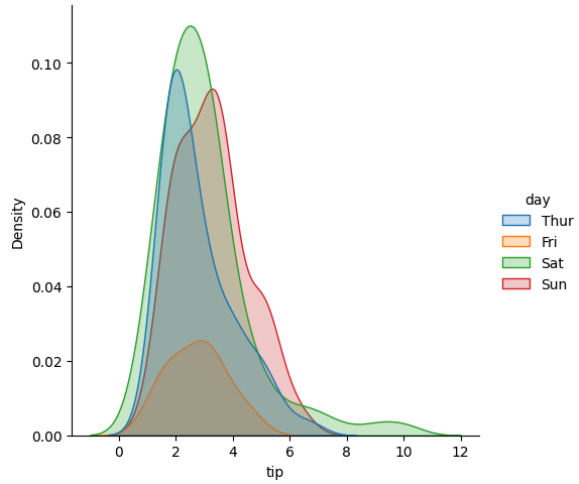
Anexo 11: Histogramas por categoría Tips



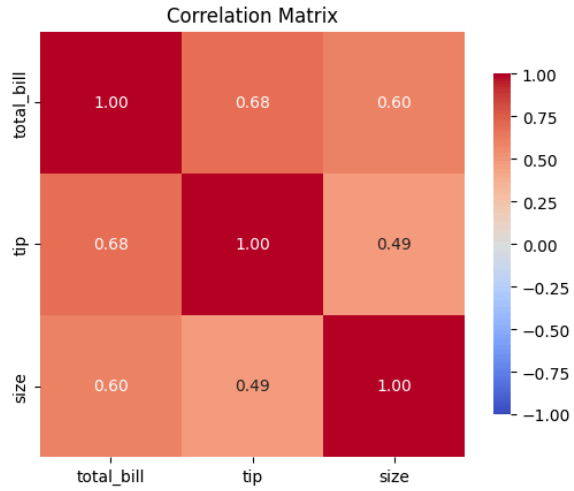
Anexo 12: Scatterplot Factura total vs Tip



Anexo 13: Pair plot en relación a tiempo Tip



Anexo 14: Gráfica de densidad por día Tip



Anexo 15: Matriz de correlación Tip