**¿Qué es y para qué sirve el análisis exploratorio de datos?**

Proceso de conocer en detalle a tus datos, darles sentido.

* Transformarlos en información útil.
* Interrogarlos para obtener las respuestas que necesites.
* Determinar cómo tratarlos.

# ¿Cómo hacer un análisis exploratorio de datos?

### ¿Por qué debería hacer un análisis exploratorio de datos?

* **Organizar y entender las variables:** podrás identificar los diferentes tipos de variables, las categorías a la que pertenecen y el tipo de análisis que puedes realizar sobre ellas.
* **Establecer relaciones entre las variables**
* **Encontrar patrones ocultos en los datos:** podrás encontrar información o comportamientos relevantes cuando hagas el EDA.
* **Ayuda a escoger el modelo correcto para la necesidad correcta:** una vez encuentres cómo están relacionadas las variables podrás descubrir las variables que más se ajustan a un tipo de modelo y de esta manera elegirás el correcto
* **Ayuda a tomar decisiones informadas:** decisiones basadas en los datos, en las relaciones que encuentres entre variables, en patrones ocultos y en los modelos que generes a través de la EDA

#### Pasos de una Análisis Exploratorio de Datos

1. **Hacer preguntas sobe los datos.** Hazte las siguientes preguntas para guiar el EDA:
   * ¿Qué te gustaría encontrar?
   * ¿Qué quisieras saber de los datos?
   * ¿Cuál es la razón para realizar el análisis?
2. **Determinar el tamaño de los datos.** Debes responder preguntas como:
   * ¿Cuántas observaciones existen?
   * ¿Cuántas variables hay?
   * ¿Necesito todas las observaciones?
   * ¿Necesito todas las variables?
3. **Categorizar las variables.** Debes preguntarte:
   * ¿Cuántas variables categóricas existen?
   * ¿Cuántas variables continuas existen?
   * ¿Como puedo explorar cada variable dependiendo de su categoría?
4. **Limpieza y validación de los datos.** En ese paso debes preguntarte:
   * ¿Tengo valores faltantes?
   * ¿Cuál es la proporción de datos faltantes?
   * ¿Como puedo tratar a los datos faltantes?
   * ¿Cuál es la distribución de los datos?
   * ¿Tengo valores atípicos?
5. **Establecer relaciones entre los datos.** Responde preguntas como:
   * ¿Existe algún tipo de relación entre mi variable X y Y?
   * ¿Qué pasa ahora si considero la variable Z en el análisis?
   * ¿Qué significa que las observaciones se agrupen?
   * ¿Qué significa el patrón que se observa?

Este proceso es cíclico. A pesar de que pueda parecer infinito, este proceso en algún momento debe salir del ciclo y continuar para obtener algún tipo de valor.

# Tipos de análisis de datos

Se dividen en 4 partes:

1. **Analítica Descriptiva:** Que sucedió? Provee de ideas sobre eventos del pasado. Permite dar una vista general de las propiedades, dimensiones y comportamiento de un fenómeno desde la mera descripción de los datos.
2. **Analítica Diagnostica:** Porque sucedió? Profundiza para encontrar las causas del evento. Es aplicar unos criterios diferenciales a los datos descriptivos para identificar una situación problemática. Permite establecer relaciones entre variables y crear hipótesis sobre las posibles causas de su comportamiento actual.
3. **Analítica Predictiva:** Que podría pasar si? Utiliza los datos del pasado para predecir un futuro evento. Utiliza datos para explicar un posible comportamiento de las variables de interés si se siguen las tendencias actuales o si estas cambian. Permite determinar si la situación problemática identificada puede representar mayor o menor riesgo en el futuro.
4. **Analítica Prescriptiva:** Que debería hacerse? Analiza decisiones y eventos del pasado para estimar la probabilidad de diferentes resultados. Utiliza datos para apoyar la toma de decisiones, con base en eventos del pasado. Propone soluciones basadas en evidencia. Igualmente, la etapa prescriptiva también puede hacer iniciar el ciclo de nuevo, ya que monitorear la solución también da pie a procesos nuevos de descripción y diagnóstico de ese monitoreo, volviendo a iniciar el ciclo una vez más.

La descriptiva y diagnostica son sobre el pasado, La predictiva y prescriptiva son sobre el futuro.

*Todas las analíticas se van construyendo una sobre la otra por lo que lleva tiempo, es decir, no puedes pasar de la Descriptiva a la Predictiva sin antes pasar por la Diagnostica*

*A medida que vas pasando de una analitica a otra tambien incrementa la complejidad ya que necesitas entender los datos a un nivel mas preciso conforme vayas avanza*

# Tipos de datos y análisis de variables

### Cualitativos

#### Categóricos

Este tipo de datos representa las características de un objeto; por ejemplo, género, estado civil, tipo de dirección o categorías de las películas. Estos datos a menudo se denominan conjuntos de datos cualitativos en estadística.

Una variable que describe datos categóricos se denomina variable categórica. Estos tipos de variables pueden tener uno de un número limitado de valores. Es más fácil para los estudiantes de informática entender los valores categóricos como tipos enumerados o enumeraciones de variables. Hay diferentes tipos de variables categóricas:

* **Ordinal**

En las escalas ordinales, el orden de los valores es un factor significativo.

Una encuesta donde se me muestran 5 valores y debo de escoger uno de ellos

* **Nominal**

Estos se practican para etiquetar variables sin ningún valor cuantitativo. Las escalas se conocen generalmente como etiquetas. Y estas escalas son mutuamente excluyentes y no tienen ninguna importancia numérica. Veamos algunos ejemplos:

Genero, los idiomas que se hablan en un país en particular, Especies biológicas, Partes de la oración en gramática, Rangos taxonómicos en biología  
Las escalas nominales se consideran escalas cualitativas y las medidas se toman utilizando las escalas cualitativas.

Ejemplo, podría ser una escala para evaluar, de cinco valores ordinales diferentes: totalmente de acuerdo / de acuerdo / neutral / en desacuerdo / totalmente en desacuerdo.  
Este tipo de escala son llamadas Likert., para este tipo de datos, se permite aplicar la mediana como medida de tendencia central; sin embargo, el promedio no esta permitido.

* **Interval**

En las escalas de intervalo, tanto el orden como las diferencias exactas entre los valores son significativos. Las escalas de intervalo se utilizan ampliamente en estadística, por ejemplo, en la medida de las tendencias centrales: media, mediana, moda y desviaciones estándar. Los ejemplos incluyen la ubicación en coordenadas cartesianas y la dirección medida en grados desde el norte magnético. La media, la mediana y la moda están permitidas en datos de intervalo.

* **Ratio**

Contienen orden, valores exactos y cero absoluto, lo que permite su uso en estadísticas descriptivas e inferenciales. Estas escalas ofrecen numerosas posibilidades para el análisis estadístico. Las operaciones matemáticas, la medida de las tendencias centrales y la medida de la dispersión y el coeficiente de variación también se pueden calcular a partir de tales escalas.

Los ejemplos incluyen una medida de energía, masa, longitud, duración, energía eléctrica, ángulo plano y volumen.

### Cuantitativos

#### Numéricos

Estos datos tienen un sentido de medición involucrado en ellos; por ejemplo, la edad, la altura, el peso, la presión arterial, la frecuencia cardíaca, la temperatura, el número de dientes, el número de huesos y el número de miembros de la familia de una persona. Estos datos a menudo se denominan datos cuantitativos en las estadísticas. El conjunto de datos numérico puede ser de tipo discreto o continuo.

* **Discreto** ⇒ Altura, peso, longitud, volumen, temperatura, humedad, edad.

Estos son datos que son contables y sus valores se pueden enumerar. Por ejemplo, si lanzamos una moneda, el número de caras en 200 lanzamientos de moneda puede tomar valores de 0 a 200 casos (finitos).

Una variable que representa un conjunto de datos discreto se denomina variable discreta. La variable discreta toma un número fijo de valores distintos. Por ejemplo, la variable País puede tener valores como Nepal, India, Noruega y Japón. La variable Rango de un alumno en un aula puede tomar valores de 1, 2, 3, 4, 5, etc.

* **Continuo** ⇒ número de amigos, calificación.

Una variable que puede tener un número infinito de valores numéricos dentro de un rango específico se clasifica como datos continuos. Una variable que describe datos continuos es una variable continua. Por ejemplo, ¿cuál es la temperatura de tu ciudad hoy?

Tipos de análisis:

* **Análisis Univariado:** analiza a cada variable por separado, entender sus característica.
* **Análisis Bivariado:** analiza la relacion de cada par de variables. Permite buscar relaciones intrinsecas entre los datos
* **Análisis Multivariado:** analiza el efecto simultaneo de multiples variables. Analiza la variables como un conjunto

*En el comienzo de la EDA, el análisis univariado te va a dar informacion mucho mas valiosa que comenzar con todas las variables al mismo tiempo. Te ayuda a analizar la contribución individual de una variable sobre las demas para luego incrementar la cantidad de variables y analizar los efectos en conjunto.*

# Herramientas de software para el análisis exploratorio de datos

Logotipo, nombre de la empresa

Descripción generada automáticamente

# Recolección de datos, limpieza y validación

*La recolección de datos es la forma de recolectar información que permite obtener conocimiento de primera mano e ideas originales sobre el problema de investigación*

**Tipos de Recolección de Datos**

* **Primaria:** Datos colectados de primera mano a través de encuestas, entrevistas, experimentos y otros. Puedes decidir que recolectas y que quiere medir. Tienes total control de los datos.
* **Secundaria:** Datos previamente recolectados por una fuente primaria externa al usuario primario. Son datos abiertos que puedes utilizar. No tienes control sobre lo que se mide o sobre el objetivo del estudio
* **Terciaria:** Datos que se adquieren de fuentes completamente externas al usuario primario. Son datos accesibles a todo el mundo por lo que pierde su valor y tienden a ser datos genéricos

**Validación de los datos**

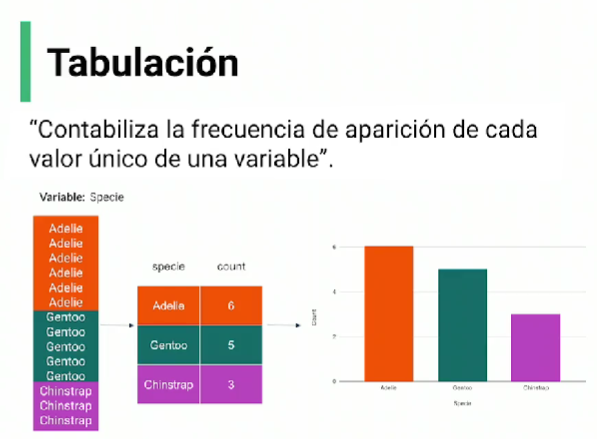
Es el proceso de asegurar la consistencia y precisión dentro de un conjunto de datos. Para asegurar la consistencia de los datos debemos validar:

* **Modelo de datos:** si un tercero recolecto los datos, verifica que preguntas quería responder con los datos. Si eres quien recolecta los datos, hazte muchas preguntas y considera si esos datos son suficientes para responderlas
* **Seguimiento de formato estándar de archivos:** verifica que la extensión de los archivos que estas manejando correspondan con el formato interno que tienen. Asegúrate que los números se expresen en el formato que estás trabajando.
* **Tipos de Datos:** verifica que los datos sen del tipo que se indica en el dataset
* **Rango de variables:** verifica que las variables estén dentro del rango establecido en la recoleccion de datos. En caso de encontrar variables fuera del rango preguntate: como llegaron esos datos aqui? tienen algun significado alterno? debo preservarlos o eliminarlos?
* **Unicidad:** verifica que tan unicos son los datos. detecta si existe duplicidad en los datos y corrige.
* **Consistencia de expresiones:** se refiere a como la persona que recolecta los datos define sus variables. Formato de fecha, de hora, variables escritas de la misma forma en toda la tabla. No son datos erroneos, solo es cuestion de darle el formato adecuado
* **Valores nulos:** pueden estar explicitos o implicitos en el dataset. Son datos faltantes. Porque esta vacio? puedo rellenarlo con otro dato? esta vacio por un proceso aleatorio o tiene un sentido?

# Explorando una variable categórica: conteos y proporciones

### Explorando una variable Categórica: Conteos y Proporciones

**Tabulacion.** Consiste en contabilizar la frecuencia de aparición de cada valor único de una variable. Los conteos y proporciones se basan en la tabulación



Estas tres formas son las más comunes de presentar un conteo: Datos ordenados, Tablas y Grafica de barras

**Proporciones.** Es la relación de correspondencia entre las partes y el todo

Gráfico, Gráfico en cascada

Descripción generada automáticamente

Tabla cruzada o tablas de contingencias: permite combinar variables categóricas

# Estadística descriptiva aplicada: medidas de tendencia central

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

***Además de la media, la mediana y la moda, existen otras medidas de tendencia central que también pueden ser útiles en diferentes contextos. Algunas de estas medidas son:***

**Media ponderada**: Para calcular la media ponderada hay que asignar un peso a los distintos valores que se van a tomar para obtener dicha media. La suma de todos los valores por sus pesos se divide finalmente por la suma de todos los pesos y se obtiene la media ponderada. se asigna un peso (importancia) a cada valor.

**Media armónica:** es el recíproco de la media aritmética. Se calcula como el número total de observaciones dividido por la suma de los recíprocos. el promedio de los reciprocos (seria 1/x donde hay un X) y no puede haber nulos.

**Media geométrica** es un tipo de media que se calcula como la raíz del producto de un conjunto de números estrictamente positivos. es la raiz n del producto de los n numeros y deben los numeros ser positivos.

#### Percentiles:

* Los percentiles dividen un conjunto de datos en 100 partes iguales.
* El percentil P indica el valor por debajo del cual caen P por ciento de las observaciones.
* Por ejemplo, el percentil 25 indica el valor por debajo del cual caen el 25 por ciento de las observaciones.

#### Cuartiles:

* Los cuartiles dividen un conjunto de datos en cuatro partes iguales.
* El primer cuartil (Q1) es el valor por debajo del cual caen el 25 por ciento de las observaciones,
* El segundo cuartil es la mediana (Q2) y
* El tercer cuartil (Q3) es el valor por debajo del cual caen el 75 por ciento de las observaciones.

### Deciles:

***Los deciles dividen un conjunto de datos en 10 partes iguales.***

* El primer decil (D1) es el valor por debajo del cual caen el 10 por ciento de las observaciones
* El segundo decil (D2) es el valor por debajo del cual caen el 20 por ciento de las observaciones, y así sucesivamente.

# Estadística descriptiva aplicada: medidas de dispersión

### Estadistica Descriptiva Aplicada: Medidas de Dispersion

* Rango: La diferencia entre el valor maximo y minimo de los datos. Da una idea de que tan dispersos estan los datos
* Rango Intercuartilico: Comprende el 25%, tanto arriba como abajo, de los datos respecto a la mediana. Divide el rango de los datos en 4 partes iguales y considera solo el 50% de los datos
* Desviacion Estandar: Ofrece la dispersion media de una variable. Si a la media de una distribucion Normal se le suma, por arriba y por debajo, la desviacion estandar se obtiene un rango que contiene el 65% de los datos. Si se suma dos desviaciones estandar se obtiene el 95% de los datos. Si se suma tres desviaciones estandar se obtiene el 99% de los datos  
  Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

  Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

#### Asimetría Estadística

Esta relacionado con la simetria de la distribucion

* Si media = mediana = moda implica que la distribucion es simetrica.
* Si media > mediana > moda, entonces La distribución esta sesgada hacia la izquierda. (Sesgo positivo)
* Si media < mediana < moda, entonces la distribucion esta sesgada hacia la derecha. (Sesgo negativo)

Diagrama

Descripción generada automáticamente

#### Cutorsis

Es un estadístico de que tan juntos o que tan dispersos están los datos respecto a la media.

* Si Cutorsis = 0, los datos estan distribuidos homogeneamente alrededor de la media (Distribucion Mesocurtica)
* Si Cutorsis > 0, los datos estan concentrados alrededor de la media (Distribucion Leptocurtica)
* Si Cutorsis < 0, los datos estan alejados de la media (Distribucion Platicurtica)

Diagrama

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

# Estadística descriptiva aplicada: distribuciones

### Estadistica Descriptiva Aplicada: Distribuciones

#### Funcion de Probabilidad de masas (PMFs)

Nos dice la probabilidad de que una variable aleatoria discreta tome un valor determinado. Ejemplo, la probabilidad de que en un salon hayan personas con una edad especifica

#### Funcion de Distribucion Acumulada (CDFs)

Devuelve la probabilidad de que una variable sea igual o menor que un valor determinado. Ejemplo, la probabilidad de que en un salon hayan personas menores o iguales a una edad especifica.

#### Función de Probabilidad de Densidad (PDFs)

Determine la probabilidad de que una variable continua tome un valor determinado. Ejemplo, la probabilidad de que en el salón hayan personar con una altura especifica

# Teorema de los grandes números

La probabilidad experimental tiende a la probabilidad teorica a medida que aumenta el numero de repeticiones del experimento. Es muy util cuando no se conoce la probabilidad teorica de un evento y tenemos la capacidad de aumentar la muestra. Mediante este teorema, al incrementar la muestra la probabilidad experimental al final se convertira en la probabilidad teorica.

#### Teorema del Limite Central

*La media de las muestras tiende aproximadamente a una distribución normal*

*La suma de n variables aleatorias independientes con medias y varianzas finitas converge en distribucion a una variable aleatoria normal*

# Estableciendo Relaciones: Graficos de puntos

Las relaciones entre dos variables puede analizarse mediante un grafico de puntos o scatterplot. Nos da una idea de la tendencia o agrupamiento de los datos. Algunas sugerencias para construir un scatterplot son:

* Cuando hay muchos puntos en la grafica no se perciben de forma correcta. Puedes mejorar la visualizacion modificando la transparencia de los puntos con el parametro alpha. Un alpha de 1/30 implica que se necesitan 30 datos para que el punto sea totalmente del color negro.
* Utilizar histogramas de dos dimensiones. Cuenta los valores que quedan dentro de un intervalo de la variable x y un intervalo de la variable ‘y’ y muestra la frecuencia utilizando una barra de colores.
* Cambio de color de los puntos. Ayuda a visualizar las zonas donde se concentra la mayor cantidad de datos

# Estableciendo relaciones: gráficos de violín y boxplots

***¿Qué pasa si tengo una variable discreta?***

Se agrega un ruido aleatorio para visualizar mejor los datos

Según los resultados de la función que dejo aquí abajo, encuentro que la especie Adelie se encuentra en las 3 islas, la especie Gentoo solo en la isla Biscoe y la especie Chinstrap solo en la isla Dream.

Los gráficos me muestran que las especies adelie y chinstrap tienen una misma distribución de pesos, mientras que la especie gentoo pesa más que el resto, lo que hace la isla Biscoe se suba el promedio de pesos, ya que solo allí están los Gentoo.

* Los gráficos de violín y los boxplots son dos tipos de gráficos utilizados en estadística y visualización de datos para representar distribuciones de datos y detectar valores atípicos.
* Un gráfico de violín es un gráfico que muestra la distribución de los datos mediante una curva simétrica que se asemeja a un violín. Este gráfico combina un histograma o un gráfico de densidad en la parte central con dos brazos que representan la densidad de los datos en ambos lados. El grosor del violín en cualquier punto representa la densidad de los datos en ese punto.
* Un boxplot, por otro lado, muestra la distribución de los datos mediante una caja con dos líneas que se extienden desde ella, conocidos como bigotes. La caja representa el rango intercuartil (IQR) de los datos, es decir, el rango entre el primer y tercer cuartil. La línea que divide la caja en dos partes iguales representa la mediana de los datos. Los bigotes se extienden hasta los valores mínimo y máximo dentro de 1,5 veces el rango intercuartil de los datos. Los valores que están más allá de los bigotes se consideran valores atípicos.

# Estableciendo relaciones: matrices de correlación

### Matrices de Correlación

La correlación expresa hasta que punto dos variables estan relacionadas entre si. Es decir, si cambian conjuntamente. Para cuantificar que tanto estan relacionadas, se utiliza el concepto de Coeficiente de Correlacion.

***Coeficiente de Correlación:****cuantifica la intensidad de la relación lineal entre dos variables en un análisis de correlación*

*La Correlación no implica causalidad. Causalidad implica “Cuando algo (la causa) genera otra cosa (efecto)*

*Diagrama, Escala de tiempo

Descripción generada automáticamente*

*Diagrama

Descripción generada automáticamente*

* Las matrices de correlación son herramientas de visualización que se utilizan para entender las relaciones entre diferentes variables en un conjunto de datos. Una matriz de correlación es una tabla que muestra las correlaciones (normalmente coeficientes de correlación de Pearson) entre todas las posibles pares de variables en un conjunto de datos.
* En una matriz de correlación, las celdas que están en la diagonal principal siempre son iguales a 1, ya que cada variable está perfectamente correlacionada con sí misma. Las celdas fuera de la diagonal principal indican el grado de correlación entre dos variables diferentes. El valor de una correlación puede variar entre -1 y 1, donde un valor de -1 indica una correlación negativa perfecta (las dos variables se mueven en direcciones opuestas), un valor de 0 indica que no hay correlación, y un valor de 1 indica una correlación positiva perfecta (las dos variables se mueven en la misma dirección).
* Las matrices de correlación son útiles para explorar la relación entre diferentes variables y pueden ser utilizadas para identificar posibles relaciones importantes para análisis posteriores. Además, las matrices de correlación pueden ser utilizadas para identificar posibles problemas de multicolinealidad en un modelo de regresión.
* Para crear una matriz de correlación en Python, podemos utilizar la biblioteca pandas

### Limitantes del Coeficiente de Correlacion

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.

* Cuando la distribucion tiene un comportamiento cuadratico el coeficiente no logra detectar esa correlacion.
* Cuando la distribucion tiene un comportamiento cubico, el coeficiente detecta la correlacion simulando que es una linea recta donde se compensa los valores por lo que hace una lectura erronea del panorama
* Siempre visualiza los datos.
* El coeficiente de correlacion no nos habla del impacto de la relacion. Un coeficiente de correlacion mas alto que otro no significa que la relacion entre las variables sea mejor o que el aporte al negocio de la correlacion mas alta sea mejor

# Estableciendo relaciones: análisis de regresión simple

### Analisis de Regresion Simple

Permite medir la fuerza del efecto en los datos mediante el ajuste de una linea recta. Es mucho mas efectivo para interpretar el comportamiento de los datos.

* Regression Lineal con scipy
* Esto arroja parametros como el slope el cual indica la pendiente de la recta (A mayor slope, mayor el efecto de la variable x sobre la variable y), intercep que indica el punto donde la linea corta el eje ‘y’ y el rvalue que nos indica la proporcion de datos capturados lo que ayuda en una mejor descripcion de los datos.
* Ejemplo donde se realiza de manera manual la regresion lineal con scipy.
* Regresion lineal con Seaborn. No da detallles sobre los datos estadisticos como scipy. Genera una zona de tendencia donde los datos de un modelo tienen mayor probabilidad de caer.
* **slope**: representa la pendiente de la línea de regresión. Indica cuánto cambia la variable dependiente por unidad de cambio en la variable independiente. En este caso, un aumento de una unidad en la longitud del pico se asocia con un aumento de 0.0119 en la masa del pingüino.
* **intercept**: representa el valor de la variable dependiente cuando la variable independiente es igual a cero. En este caso, el intercepto indica la masa del pingüino cuando la longitud del pico es cero, lo que no tiene mucho sentido en términos prácticos.
* **rvalue**: representa el coeficiente de correlación entre las dos variables. Indica cuánto se ajustan los datos a la línea de regresión. El valor de **rvalue** oscila entre -1 y 1, y un valor cercano a 1 indica una correlación positiva fuerte entre las variables. En este caso, el valor de **rvalue** es 0.6416, lo que indica una correlación moderadamente fuerte entre la longitud del pico y la masa del pingüino.
* **pvalue**: representa el valor p asociado a la hipótesis nula de que no hay relación entre las dos variables. Un valor p pequeño (generalmente < 0.05) indica que la hipótesis nula puede ser rechazada y que hay evidencia suficiente para decir que existe una relación significativa entre las variables. En este caso, el valor de **pvalue** es 3.7481e-09, que es muy pequeño y sugiere que hay una relación significativa entre la longitud del pico y la masa del pingüino.
* **stderr** y **intercept\_stderr**: representan el error estándar de la pendiente y el intercepto, respectivamente. Indican cuánto varía la pendiente y el intercepto en diferentes muestras de datos. En este caso, valores más pequeños indican que el ajuste del modelo es más preciso y confiable.

# Limitaciones del análisis de regresión simple

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

# Análisis de regresión múltiple

Permite establecer relacion entre multiples variables. Es capaz de detectar comportamientos no lineales. Ademas, permite deducir si la relacion entre las variables puede ser la causa del comportamiento de la variable objetivo.

# Visualización del análisis de regresión múltiple

### Visualizacion del Analisis de Regresion Multiple

Para tener una comparacion optima del comportamiento de cada modelo con el comportamiento de los datos, primero creamos una tabla de comparacion con la prediccion de cada modelo y los datos reales.

* Creacion Tabla de Resultados

La distribucion que se muestra en values\_result esta un poco sesgada. Ahora ubicamos el modelo que mas concuerda con la distribución de value\_result. A simple vista notamos que la distribucion que mas se asemeja es la del modelo 5

* Grafica de los resultados de la tabla mediante la Funcion de Distribucion Acumulada (ECDFs)
* Grafica de solo el comportamiento de los datos reales y el modelo que mas se asemeja

Debemos asegurarnos que el modelo no este simulando el ruido en los datos sino la cualidad que se desea.

* Grafica de los resultados de la tabla mediante la Densidad de Probabilidad

*Pudimos haber encontrado las mismas conclusiones del modelo 5 si en el Analisis Exporatorio hubieramos experimentado con un grafico entre la longitud de las aletas y el peso de los pinguinos utilizando el sexo como variable categorica. El analisis exploratorio de los datos es sumamente importante ya que nos ayuda a intuir el modelo desde el inicio y de esa manera a ahorrar tiempo.*

# Análisis de regresión logística

*l parametro P > abs(Z) indica cuales son las variables mas significativas estadisticamente para el resultado. Mientras mas tienda al valor 0 mas significativa es.*

*Cuando tenemos tres variables y en los resultados de la regresion logistica solo aparecen dos, indica que los esos valores son respecto a la variable faltante*

# Paradoja de Simpson

*Fenomeno en el cual es posible concluir dos cosas totalmente opuestas a partir de los mismos datos, dependiendo el como se clasifican estos*

Visualiza siempre los datos para tener una mejor idea de las variables que estan relacionadas.

Utiliza pairplot de la libreria Seaborn para obtener un resumen visual de las graficas de par de variables y tener una idea desde el inicio de la EDA sobre la relacion de las variables

# ¿Qué hacer cuando tengo muchas variables?

Cuando se tiene muchas variables, un analisis de pares de variables puede ser confuso por lo que tenemos que recurrir a tecnicas que nos ayudan a entender la variacion de todos los datos de manera simple: Reduciendo las dimensiones para obtener un unico espacio (Pasar de 10 variables a solo 2). Algunas de estas tecnicas son:

* Analisis de Componentes Principales (PCA): un ejemplo de utilidad es la demostracion de que los genes reflejan la geografia de Europa
* TSNE (T - Distributed Stochastic Neighbor Embedding): Separacion de todos los tipos de cancer
* UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction): intenta capturar la estructura global preservando la estructura local de los datos utlizando proyecciones en un plano
* Comparacion: algoritmo de reduccion de dimension vs conjunto de datos

# Diversidad de gráficas al explorar datos

Preguntate, Que quieres mostrar? Una distribucion, una comparacion, una relacion o una composicion?

*Si bien existen reglas que ayudaran a encontrar el grafico apropiado para tu problema, tu imaginacion siempre debe estar abierta a crear e iterar sobre el aspecto y comunicacion de los graficos*

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

Diagrama

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente