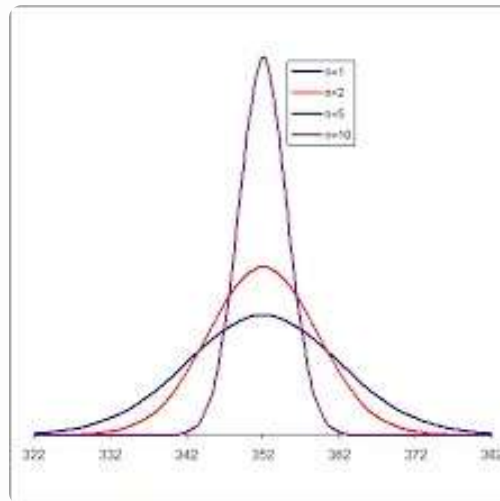


Análisis de Datos: Más Allá de la Estadística Descriptiva

El análisis de datos es una parte esencial en muchos campos. Puedes pensar en ello como la ciencia y el arte de tomar un montón de números y convertirlos en información que podamos entender y usar.

FÓRMULAS (DATOS NO AGRUPADOS)		
Media Aritmética	Mediana	Moda
$\bar{x}$	$\tilde{x}, Me, x_{me}$	Mo, x_{mo}
$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$	Cuando n es impar $Me = \frac{x_{n+1}}{2}$	Mo = El valor que mas se repite
Sumatoria de los valores observados divididos entre su cantidad	Cuando n es par $Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$	

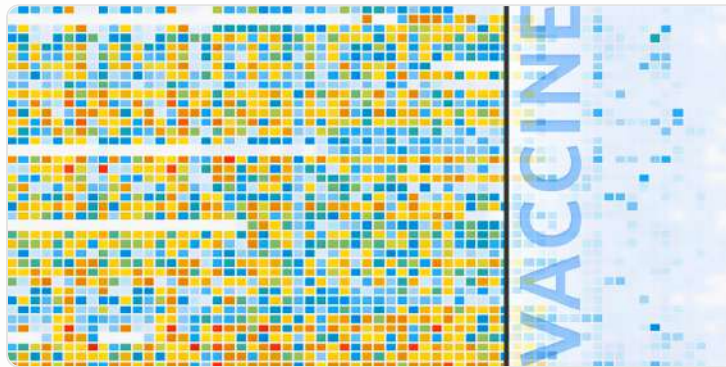
La estadística descriptiva es nuestra primera herramienta en análisis de datos. Nos ofrece un resumen rápido y sencillo de los datos, proporcionándonos un vistazo inicial sobre posibles patrones y tendencias.



El cuarteto de Anscombe: las estadísticas simples pueden ser las mismas para diferentes conjuntos de datos, y por eso es muy importante siempre mirar los gráficos de tus datos.

Las estadísticas descriptivas, aunque útiles, no cuentan toda la historia. Pueden ser engañosas al ocultar la variación real y las distribuciones de los datos. Por lo tanto, es esencial explorar los datos más a fondo.

Ejemplo de visualizaciones

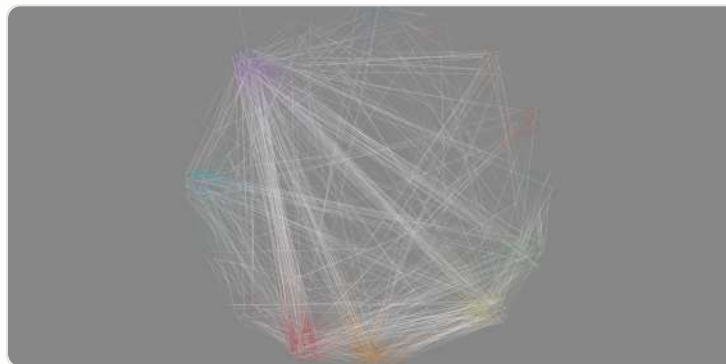


WSJ WSJ



Battling Infectious Diseases in the 20th Century: The Impact of Vac...

The number of infected people, measured over 70-some years and across all 50 states and the District of Columbia, generally declined after vaccines we...



F FlowingData



A Day in the Life of Americans

I wanted to see how daily patterns emerge at the individual level and how a person's entire day plays out. So I simulated 1,000 of them.

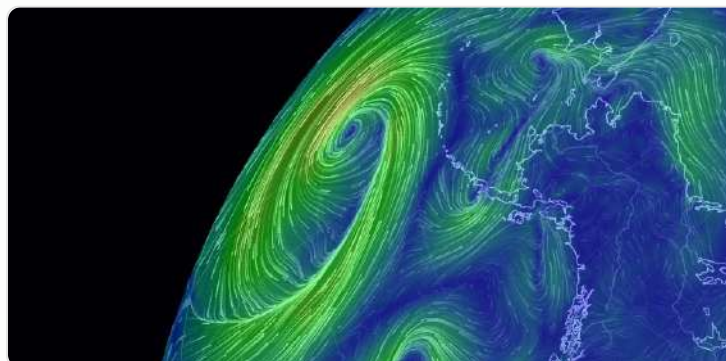


Information is Beautiful



Information is Beautiful

Distilling the world's data, information & knowledge into beautiful infographics & visualizations



earth.nullschool.net



earth :: a global map of wind, weather, and ocean conditions

See current wind, weather, ocean, and pollution conditions, as forecast by supercomputers, on an interactive animated map. Updated every three hours.

El cuarteto de Anscombe



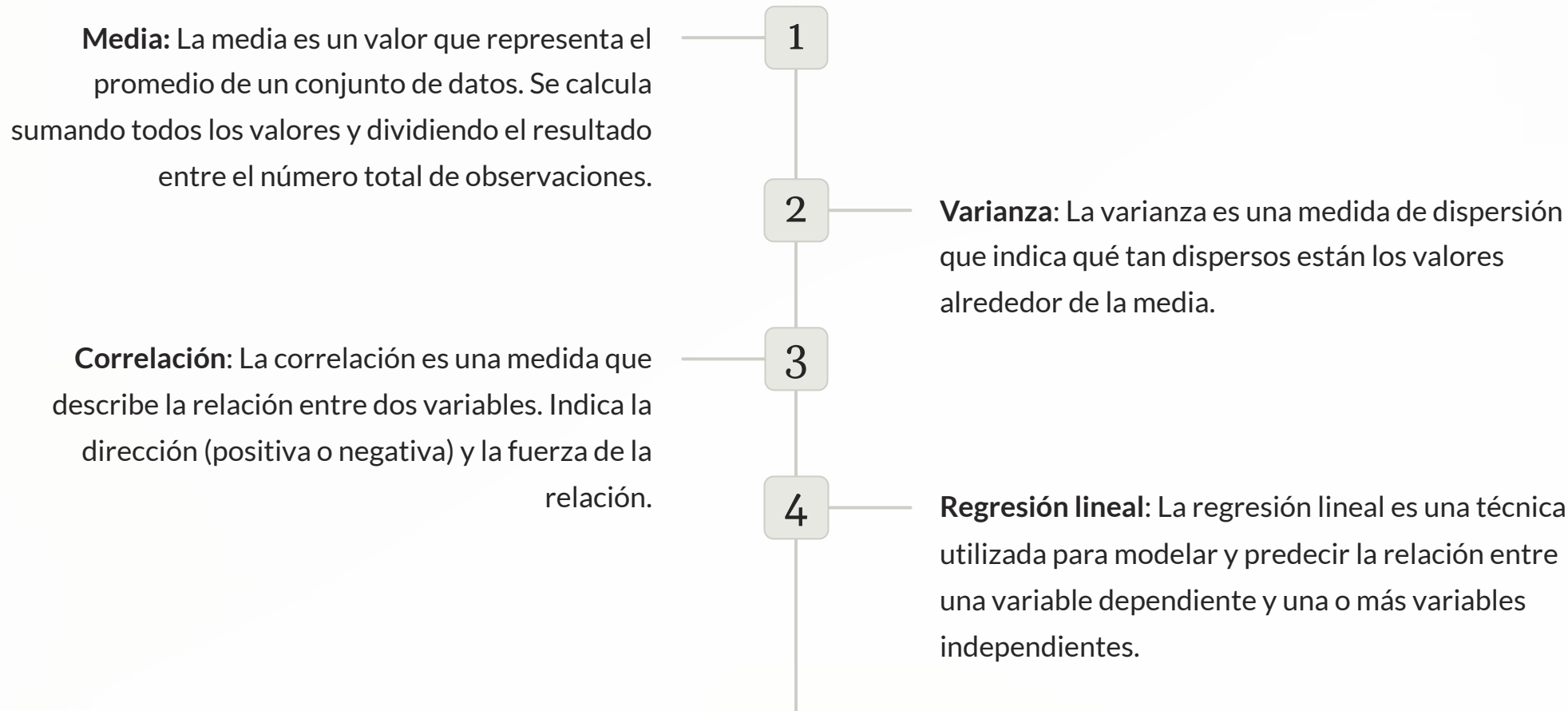
Francis John "Frank" Anscombe, quien nació el 13 de mayo de 1918 y falleció el 17 de octubre de 2001. Fue un influyente estadístico y profesor universitario, conocido por sus contribuciones en el campo de la estadística y el análisis de datos.

El Cuarteto de Anscombe es un conjunto de cuatro conjuntos de datos.

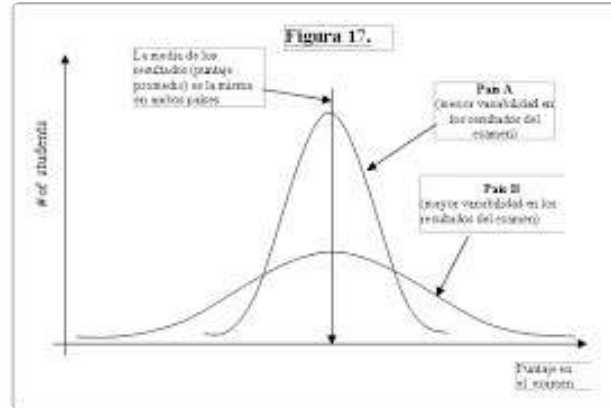
Cada uno con 11 puntos, que tienen propiedades estadísticas idénticas o muy similares.

A pesar de que los datos son distintos en su forma gráfica.

Conceptos Fundamentales de Estadística



Comparación de Descriptores Estadísticos en el Cuarteto de Anscombe



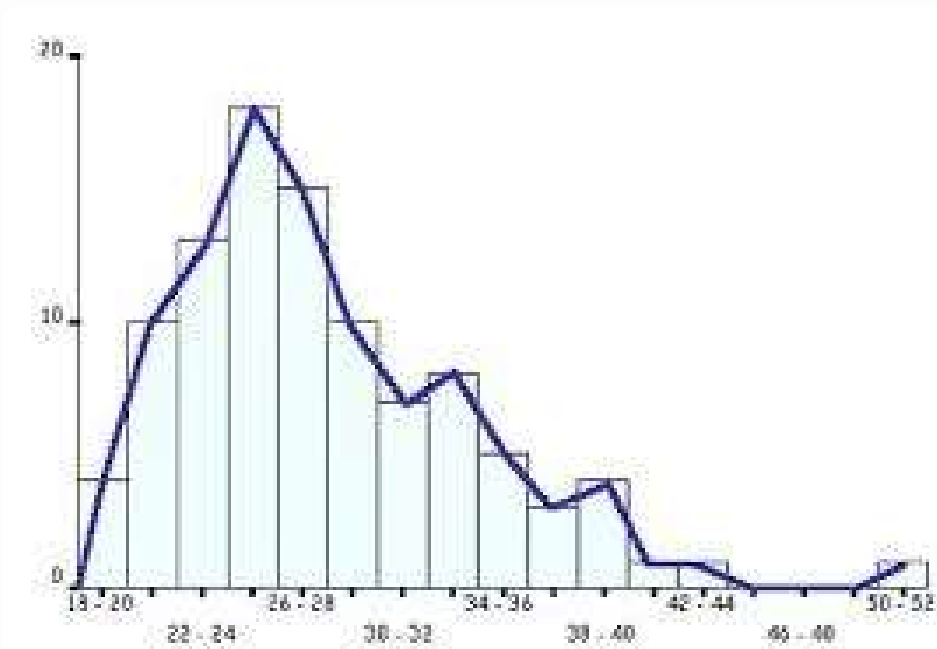
Media: Los valores medios (promedios) de las coordenadas x e y en cada conjunto de datos son aproximadamente iguales.

Varianza: Las varianzas de las coordenadas x e y en cada conjunto de datos son aproximadamente iguales.

Correlación: Los coeficientes de correlación entre las coordenadas x e y en cada conjunto de datos son aproximadamente iguales.

Regresión lineal: Las líneas de regresión lineal ajustadas a los datos en cada conjunto tienen la misma ecuación.

¿Qué es un histograma?

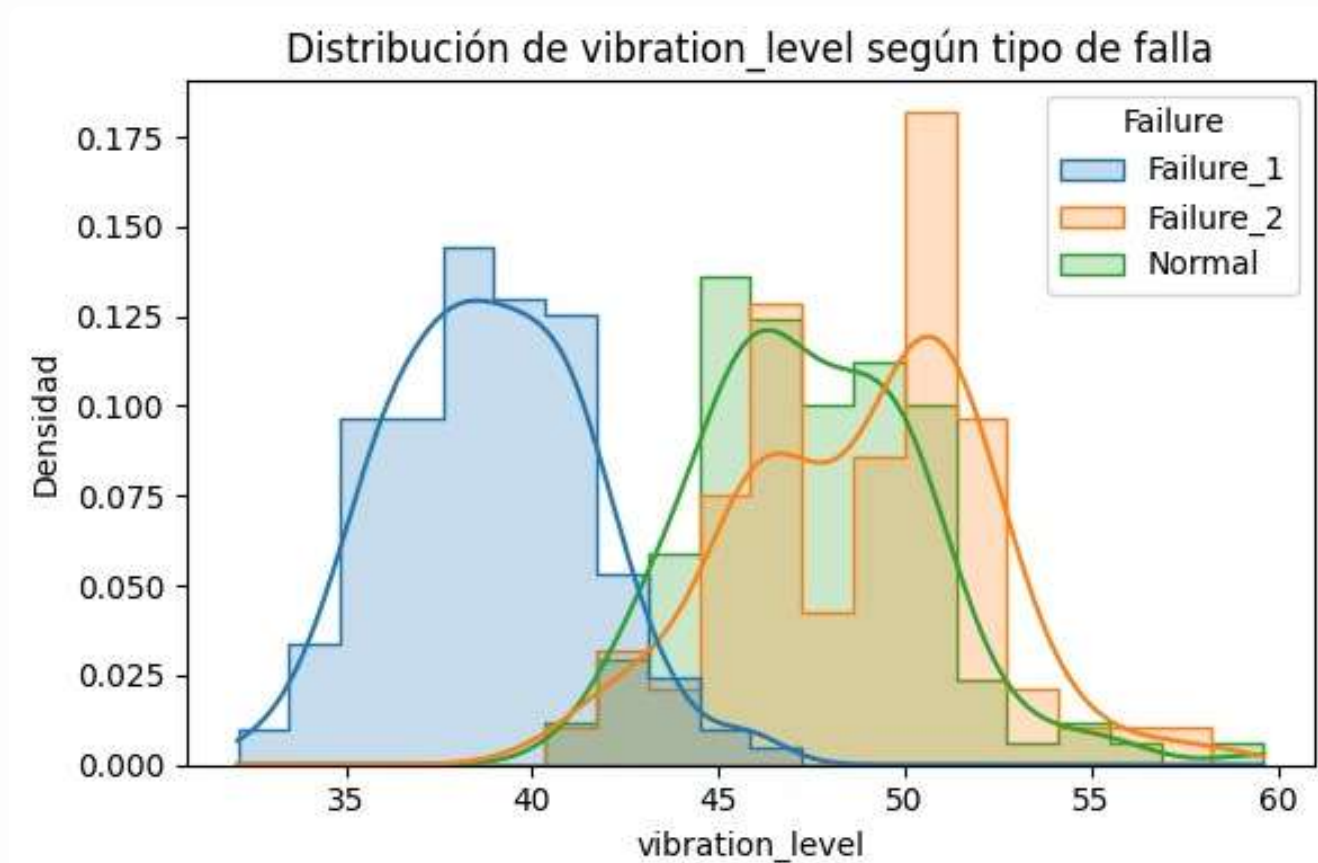


Un histograma es una representación gráfica de la distribución de un conjunto de datos numéricos. Los datos se dividen en intervalos o "bins" y se muestra el número de datos en cada uno de estos intervalos. En otras palabras, un histograma proporciona una visualización de la frecuencia de los datos.

Un histograma estándar consta de los siguientes componentes:

- **Bins:** Estos son los intervalos en los que se dividen los datos. Cada bin tiene un rango mínimo y máximo, y todos los valores de los datos que caen dentro de este rango se cuentan para ese bin.
- **Altura de las barras:** La altura de cada barra en un histograma indica la cantidad de datos que caen dentro de ese bin en particular. Si la altura de la barra es alta, significa que muchos datos caen dentro de ese rango, y si la altura es baja, significa que pocos datos caen dentro de ese rango.

¿Por qué usar un histograma?

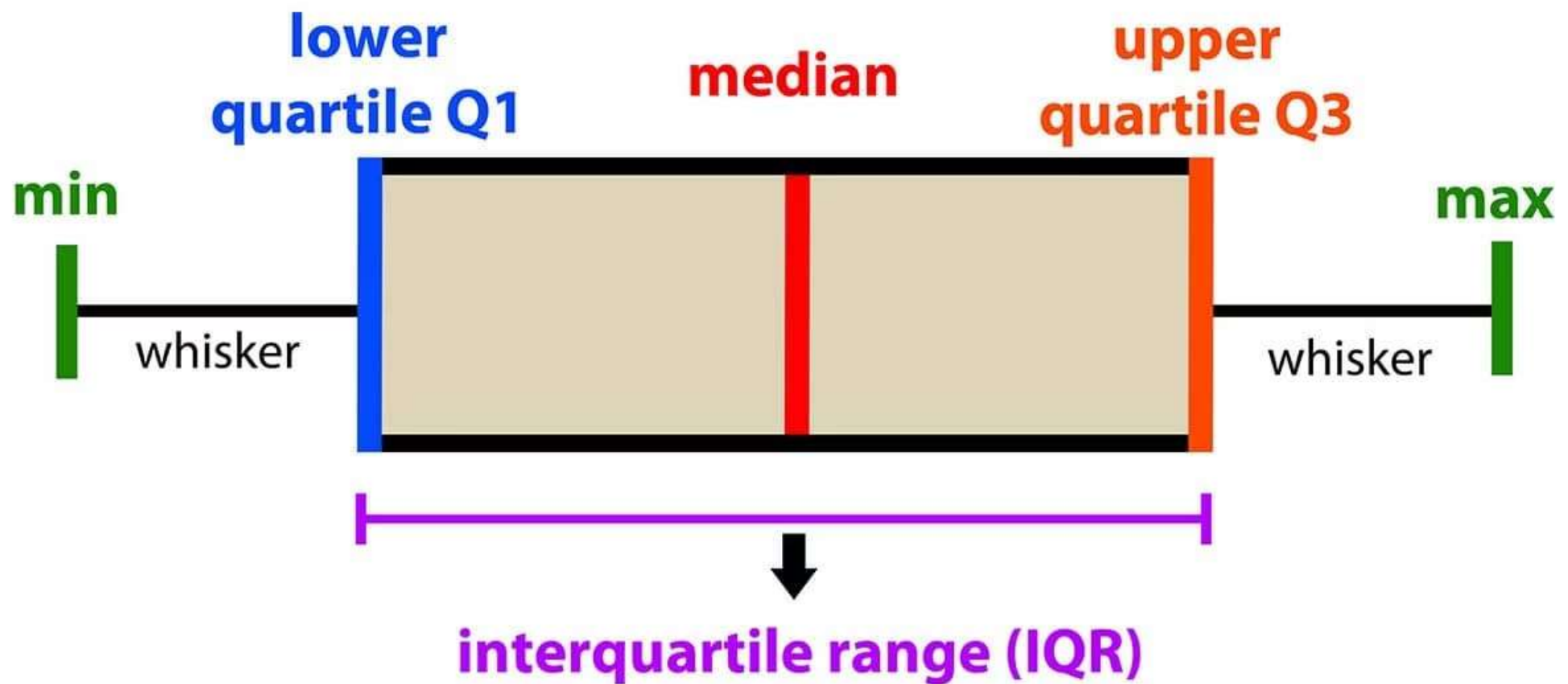


El histograma es una herramienta útil para la visualización de datos por varias razones:

- **Visualización de la distribución de los datos:** Un histograma proporciona una visualización de la forma general de la distribución de los datos. Puedes ver rápidamente dónde se centran los datos, si los datos están sesgados a la izquierda o a la derecha, o si los datos están uniformemente distribuidos.
- **Identificación de anomalías:** Los histogramas también son útiles para identificar cualquier anomalía en los datos. Por ejemplo, si un histograma tiene dos picos (una distribución bimodal), puede indicar que tienes dos grupos distintos de datos.
- **Comparación de distribuciones:** Si tienes varios histogramas en la misma escala, puedes comparar fácilmente la distribución de diferentes conjuntos de datos.

Que es un Box plot

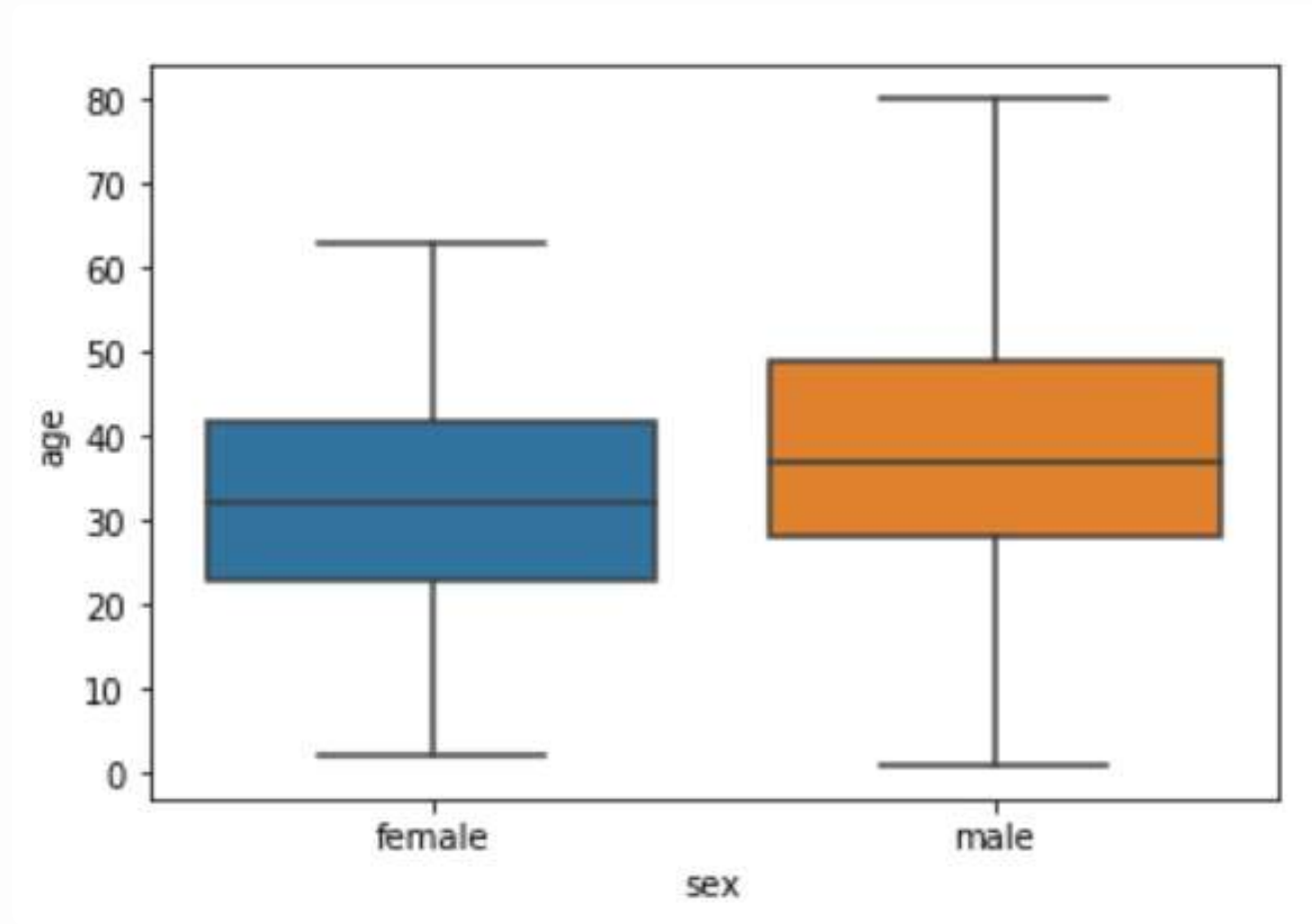
introduction to data analysis: Box Plot



Un boxplot estándar consta de los siguientes componentes:

- **Caja:** La caja en el medio del gráfico representa el rango intercuartílico (IQR), que es la distancia entre el primer cuartil (Q1, el valor que queda por encima del 25% inferior de los datos) y el tercer cuartil (Q3, el valor que queda por encima del 25% superior de los datos). La línea en medio de la caja representa la mediana (o segundo cuartil), que es el valor medio del conjunto de datos.
- **Bigotes:** Los bigotes se extienden desde la caja hasta el valor más pequeño y más grande dentro del rango de $1.5 * \text{IQR}$ desde la caja. Este rango se considera el rango "normal" o esperado de los datos.
- **Puntos:** Cualquier punto que quede fuera de los bigotes se considera un valor atípico o anómalo.

Por qué usar un boxplot?



El boxplot es una herramienta de visualización de datos muy útil por varias razones:

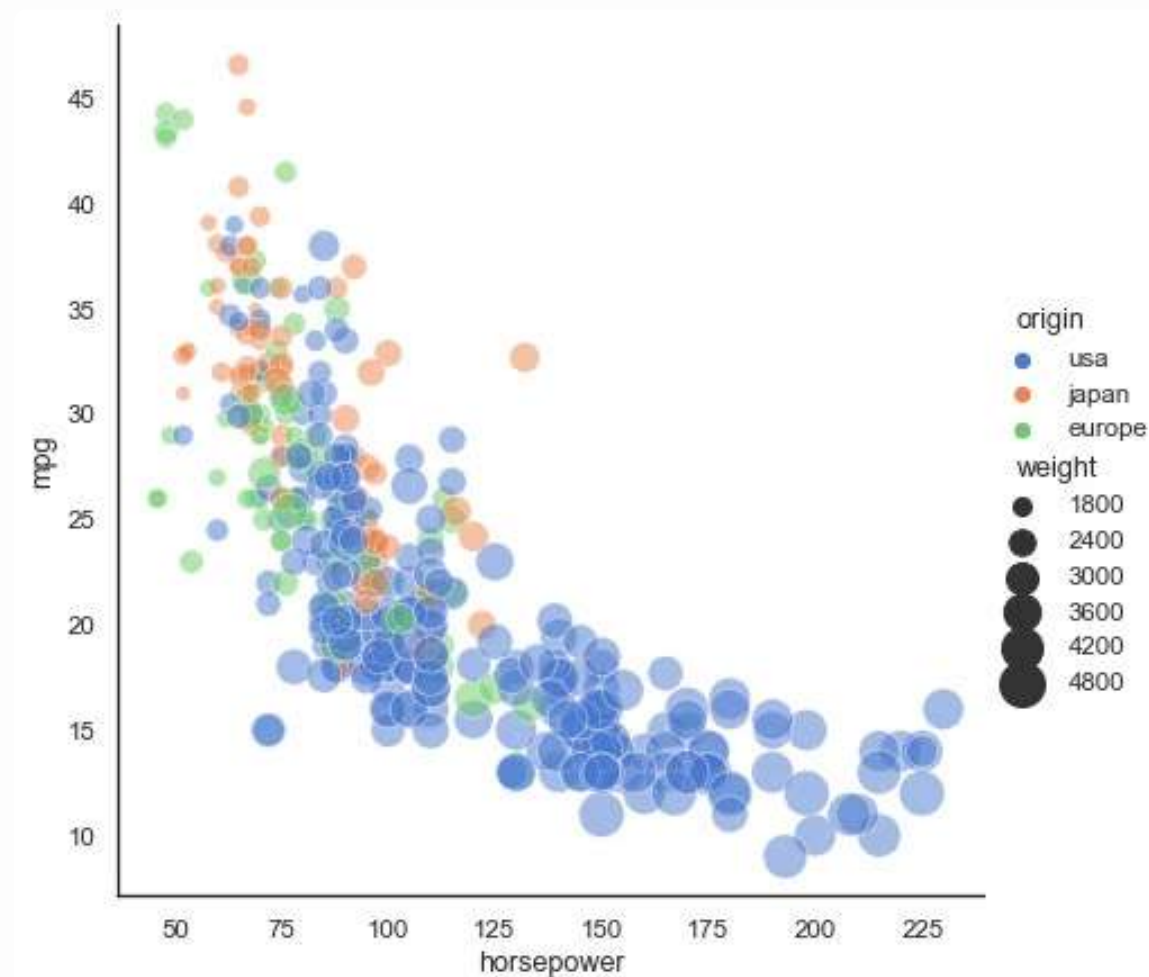
- **Resumen de los datos:** Un boxplot proporciona un resumen visual de la distribución de un conjunto de datos en un solo gráfico. En lugar de tener que mirar varias estadísticas diferentes (como la media, la mediana, los cuartiles, etc.), puedes ver todas estas medidas en un solo gráfico.
- **Identificación de valores atípicos:** Los boxplots son una excelente manera de identificar valores atípicos en los datos. Cualquier punto que se muestre fuera de los bigotes se considera un valor atípico.
- **Comparación de distribuciones:** Los boxplots son especialmente útiles cuando se comparan las distribuciones de varios conjuntos de datos.

Introducción a Scatter plots



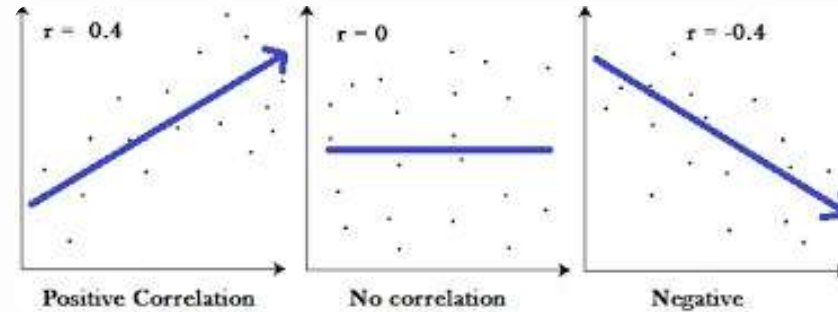
1. **Análisis de correlación:** Los scatter plots son extremadamente útiles para visualizar y entender las correlaciones entre dos variables numéricas. Por ejemplo, puedes usar un scatter plot para ver si existe una relación lineal o no lineal entre las variables.
2. **Detección de patrones y tendencias:** Los gráficos de dispersión pueden ayudar a detectar tendencias, patrones o anomalías en los datos. Esto es particularmente útil cuando tienes una gran cantidad de datos, donde las tendencias pueden no ser inmediatamente evidentes.
3. **Identificar outliers:** Los scatter plots también son útiles para identificar los outliers en tu conjunto de datos. Un outlier será un punto que esté distante del resto de los puntos en el gráfico.

Cuándo usar un Scatter Plot



1. **Cuando ambas variables son numéricas:** Los scatter plots solo pueden ser utilizados cuando ambas variables son numéricas. No son apropiados para variables categóricas.
2. **Cuando quieres entender la relación entre dos variables:** Si estás interesado en entender si existe una relación entre dos variables y cómo se comportan entre sí, un scatter plot es una buena opción.
3. **Cuando tienes una gran cantidad de datos:** Los scatter plots pueden ser particularmente útiles cuando tienes un gran conjunto de datos, ya que pueden mostrar patrones que no serían visibles en un formato tabular.

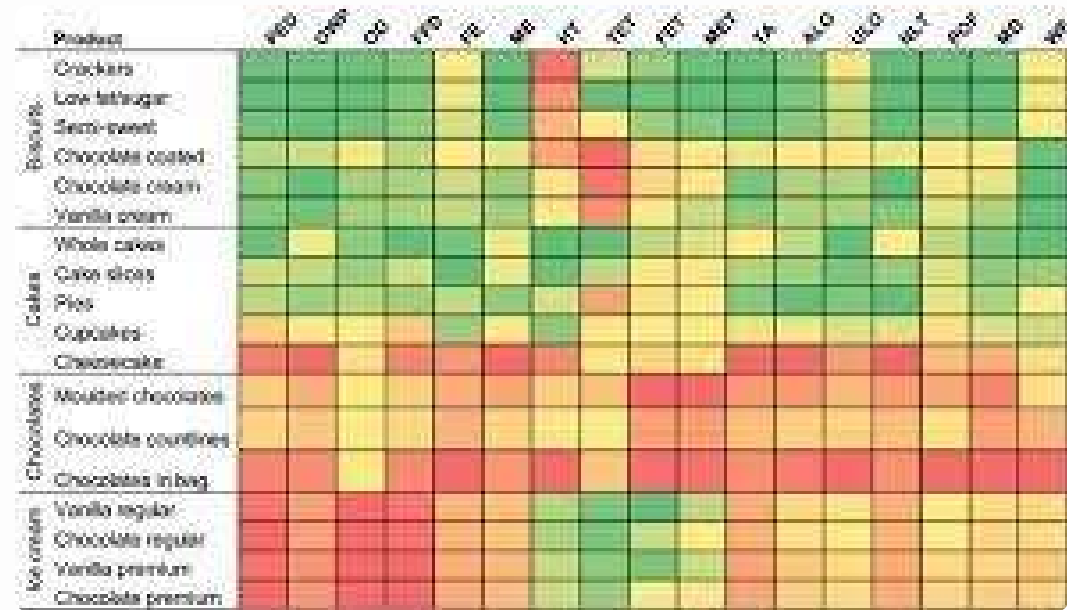
Análisis de correlación



El análisis de correlación es una técnica estadística utilizada para medir el grado de relación entre dos variables. Los coeficientes de correlación oscilan entre -1 y +1, donde +1 indica una correlación perfecta positiva, -1 una correlación perfecta negativa, y 0 indica que no hay ninguna relación.

En Python, puedes usar el método `corr()` de pandas para calcular la matriz de correlación, y luego usar la biblioteca seaborn o matplotlib para visualizarla.

Heatmaps



Los mapas de calor (Heatmaps) son una gran herramienta para visualizar la correlación entre múltiples variables al mismo tiempo. Cada celda en el gráfico está coloreada según el valor de correlación entre las variables correspondientes.

Cuándo utilizar Heatmaps

1. **Para visualizar una matriz de correlación:** Como se mencionó anteriormente, los mapas de calor son una forma excelente de visualizar una matriz de correlación y, por lo tanto, las relaciones entre múltiples variables.
2. **Para visualizar patrones complejos:** Los mapas de calor pueden ser particularmente útiles cuando se trabaja con datos complejos donde los patrones pueden no ser inmediatamente evidentes.
3. **Para visualizar variables categóricas:** Aunque los scatter plots son excelentes para visualizar relaciones entre variables numéricas, los mapas de calor pueden ser útiles para visualizar relaciones entre variables categóricas.

El dataset de pingüinos

1

Estadística descriptiva:

Usemos un script de python. Que podemos concluir?, podemos agrupar por otro concepto?

2

Histogramas: Queremos mezclar variables categoricas y continuas.

3

Boxplots: Nuevamente queremos mirar una variable continua pero separarla por variable categorica.

4

Scatter plot: Excelente para representar la relación entre dos variables continuas. Pueden ser enriquecidos al incorporar variables categóricas en la visualización.