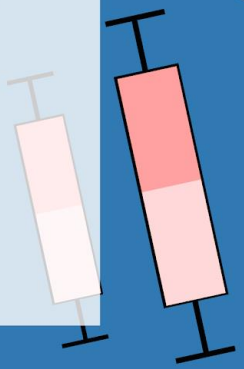
Herramientas Estadísticas y Forecast

Clase 2 – Análisis descriptivo y gráfico

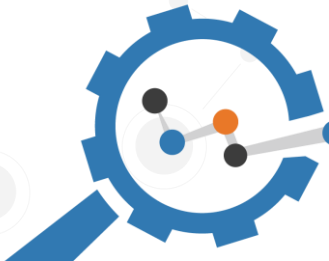


DATA SCIENCE UC
FACULTAD DE MATEMÁTICAS

FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

Contenido del curso

1. Introducción a la estadística y análisis descriptivo
2. Análisis descriptivo y gráfico
3. Probabilidad y Distribuciones
4. Muestreo
5. Inferencia estadística: Pruebas de hipótesis
6. Taller Práctico de inferencia
7. Introducción a los modelos estadísticos
8. Modelos predictivos I: Modelos de regresión lineal.
9. Modelos predictivos II: Regresión logística y otros modelos.
10. Modelos de Forecasting I.
11. Modelos de Forecasting II
12. Taller de Forecast



Actividad

Se posee un extracto de la Encuesta Nacional de Salud (ENS) con algunas características de interés, que son útiles para realizar pronósticos en la población chilena.

I. Explore los datos.

II. Clasifique todas las variables según la naturaleza de los datos.

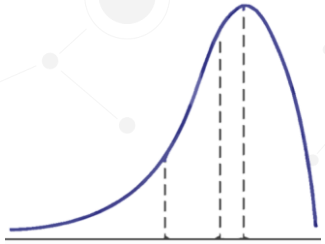
III. Obtenga una tabla de contingencia para estudiar el efecto de la diabetes en relación con si fuma o no fuma

IV. Calcule la variable $\text{IMC} = \frac{Peso}{Talla\ en\ metros^2}$, luego categorícela como: $normal < 25,25 \leq sobre\ peso \leq 30,\ obeso > 20$

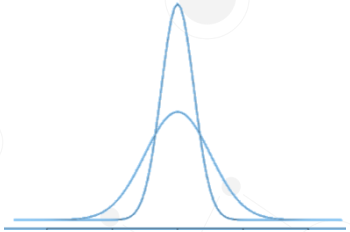
V. Obtenga una tabla de contingencia entre el IMC categorizada y diabetes, ¿qué puede concluir?

Indicadores Descriptivos

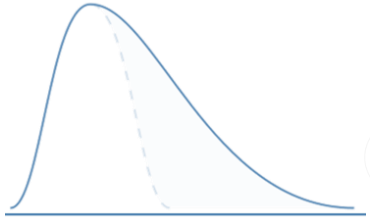
Al tener datos numéricos, se puede obtener indicadores que intenten representar a todos los posibles resultados de la variable. En efecto, para poder realizar dicha representación, necesitaremos construir medidas asociadas a:



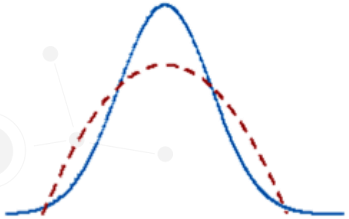
Posición



Dispersión



Asimetría



Curtosis

Indicadores Descriptivos

- Para poder representar los datos a través de su posición, las medidas más frecuentes son

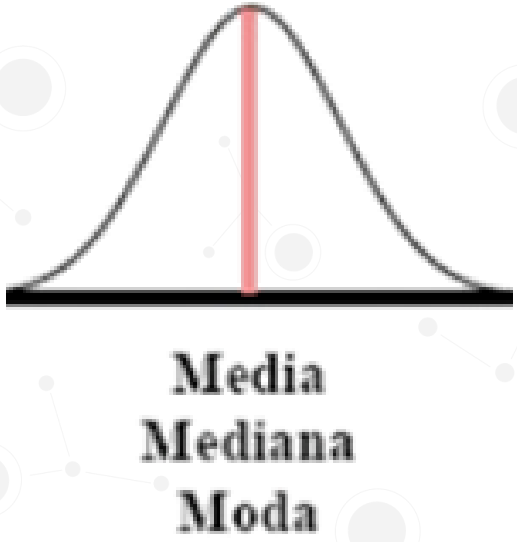
- La **Media** es el indicador de centralidad que considera la el valor y frecuencia de datos para buscar el mejor representante. En estadística, la más común es la **Media Ponderada**, o **Valor Esperado**.
  -

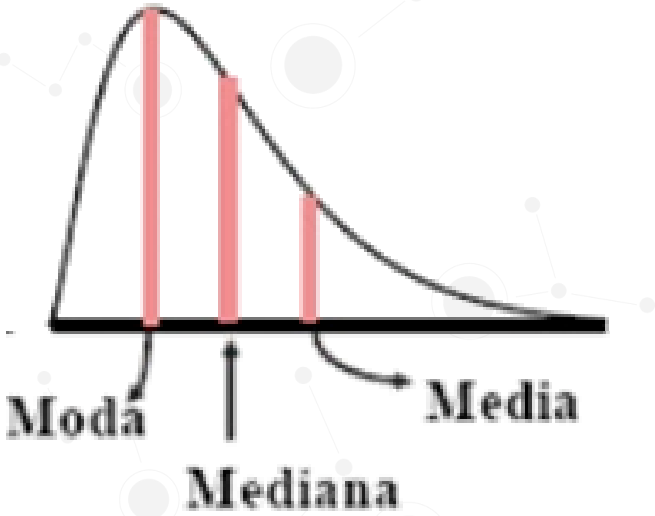
- La **Mediana** corresponde a la ubicación del dato del medio, o bien, donde se ubica a lo más el 50% de los datos.

- La **Moda** es el o los valores que poseen más frecuencia en una observación.

Indicadores Descriptivos

Entre las más usuales…¿Cuándo es mejor usar cada una?





Indicadores Descriptivos

Uno de los factores que influyen en la decisión es el tipo de datos: categóricos o numéricos.

- Si los datos son **numéricos** tiende a preferirse la media en casos simétricos y la mediana en casos asimétricos.

- Si los datos son **categóricos ordinales** es más apropiada la mediana o la moda.

- Si los datos son **categóricos nominales**, el más apropiado es la moda.

Indicadores Descriptivos

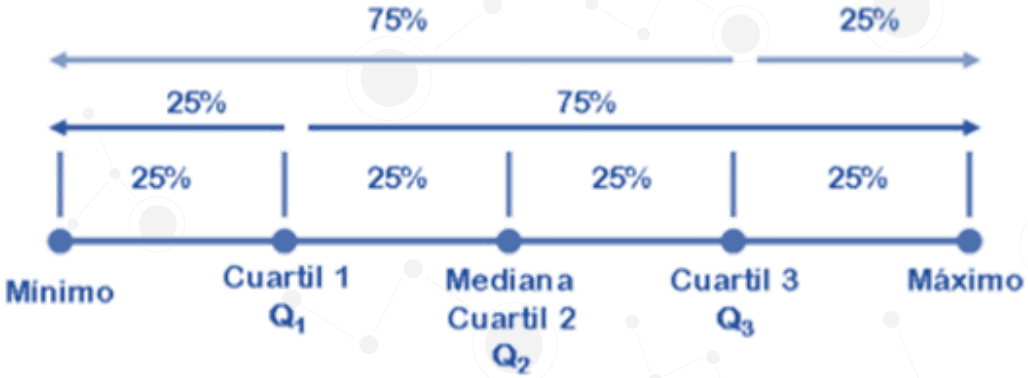
Un **cuantil** corresponde a la ubicación del dato según ciertas divisiones:

Los **cuartiles** son divisiones de los datos en cuatro partes, tal que el primer cuartil representa hasta el primer 25%, el segundo cuartil corresponde al 50% (igual a la mediana), y el tercer cuartil es la ubicación del dato en la posición 75%.

También podemos encontrar **quintiles**, al dividir en 5 grupos, **deciles** al dividir el 10 grupos, y **percentiles** al dividir en 100.

Indicadores Descriptivos

Suponga que tiene todos los datos ordenados de menor a mayor. Entonces, se puede identificar los cuartiles de la siguiente forma:



Indicadores Descriptivos

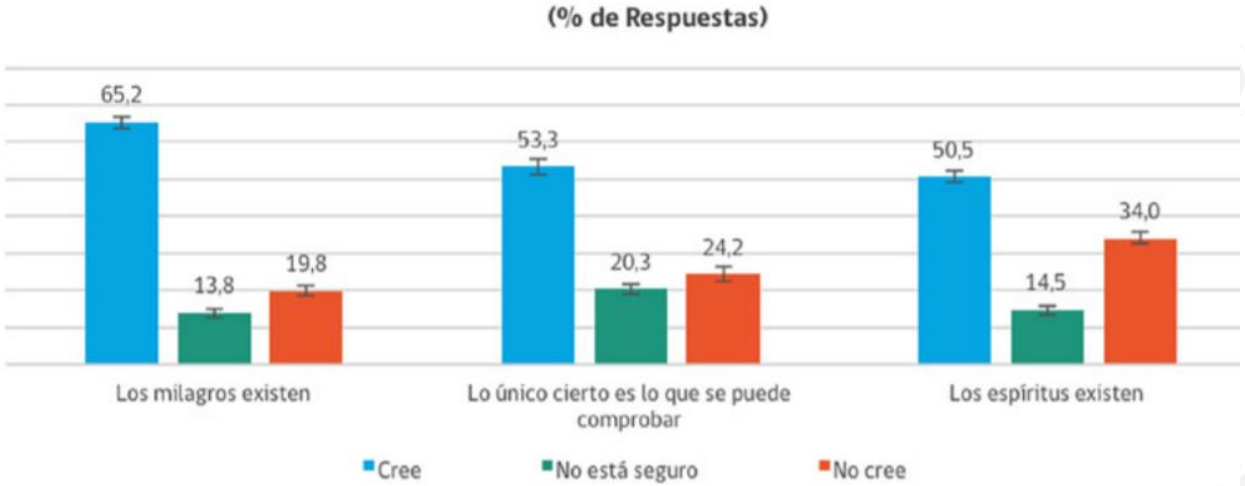
Suponga el tiempo de respuesta de un mensaje por WhatsApp

Mínimo	0
Primer Cuartil	1.43
Media	4.87
Mediana	3.51
Moda	0
Tercer Cuartil	6.61
Máximo	34.56

Indicadores Descriptivos

Un clásico "problema"

Gráfico 12. P1. ¿Me podría decir si usted cree, no está seguro de creer o no cree? – Totales.



Cree	1
No está seguro	2
No Cree	3

El promedio de "los espíritus existen" es 1.8 ≈ 2 → El promedio es "no está seguro"

Indicadores Descriptivos

- Las indicadores de dispersión miden la variabilidad de las observaciones. Entre las más usuales encontramos:
- El **Rango** corresponde a la diferencia del mínimo y el máximo.
- El **Rango Inter cuartil (RIC)** es la distancia donde se concentra el 50% central, entre el primer y tercer cuartil.
- La **Varianza** es un indicador sin interpretación, que sirve para calcular la **Desviación Estándar**, que corresponde a la dispersión de datos respecto a la media.
- El **Coeficiente de Variación** es adimensional, y responde a la relación de desviación vs media.

Indicadores Descriptivos

¿Cuándo usar cada una?

- Si los datos son **numéricos simétricos** se recomienda la desviación estándar y el coeficiente de variación.
- Si los datos son **numéricos** asimétricos, es mejor usar el rango y el rango Inter cuartil.
- Si los datos son **categóricos ordinales** es más apropiado un rango

Indicadores Descriptivos

Suponga el tiempo de respuesta de un mensaje por WhatsApp

Rango	35
RIC	5.17
Varianza	24.25
Desv. Estándar	4.92
Coef. Variación	1

Indicadores Descriptivos

Otros indicadores que le van dando forma a los datos se relacionan con el **coeficiente de asimetría**, que cuantifica el nivel asimétrico entre los datos, mientras que la **curtosis** cuantifica la concentración en el centro.

En ambos casos, mientras más cercano a 0 indica la estabilidad.

Indicadores Descriptivos

Suponga el tiempo de respuesta de un mensaje por WhatsApp

Asimetría	2.01
Curtosis	2.28

Actividad

A partir de los datos de la Encuesta nacional de salud,

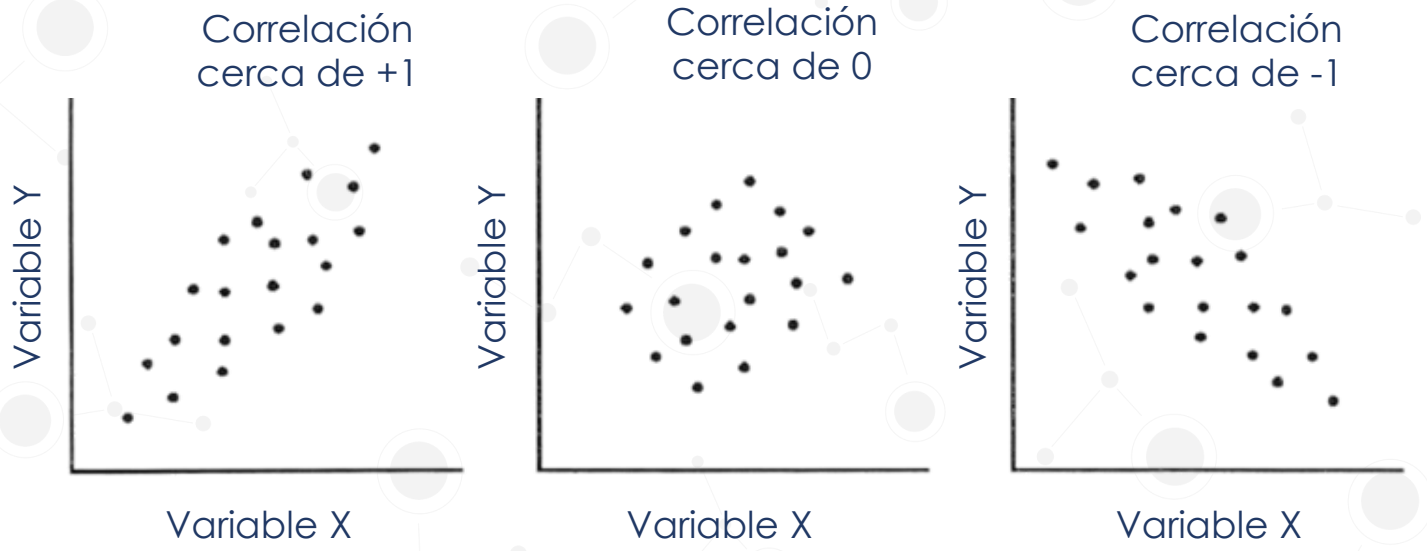
I. Obtenga un resumen numérico completo del colesterol.

II. Obtenga un resumen numérico completo del peso.

III. Obtenga un resumen numérico completo del HDL, por genero.

Indicadores Descriptivos

Para medir la relación entre dos variables, se obtiene la **covarianza**, que está relacionada a la dispersión de una variable respecto a la otra. Sin embargo, la interpretación corresponde a la **correlación**, que mide la relación lineal entre dos variables, tomando valores entre -1 y 1.



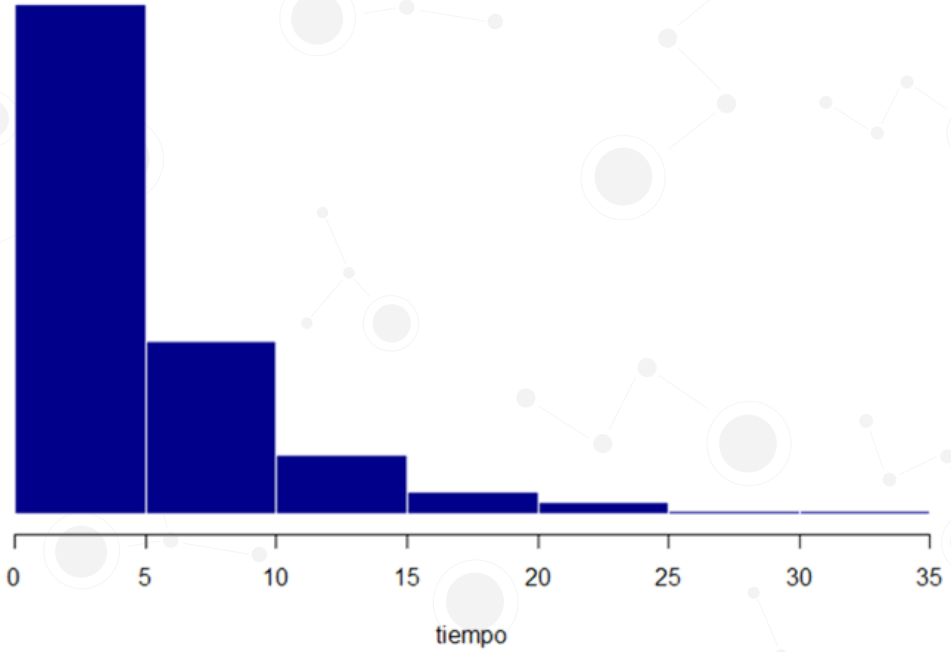
Indicadores Gráficos

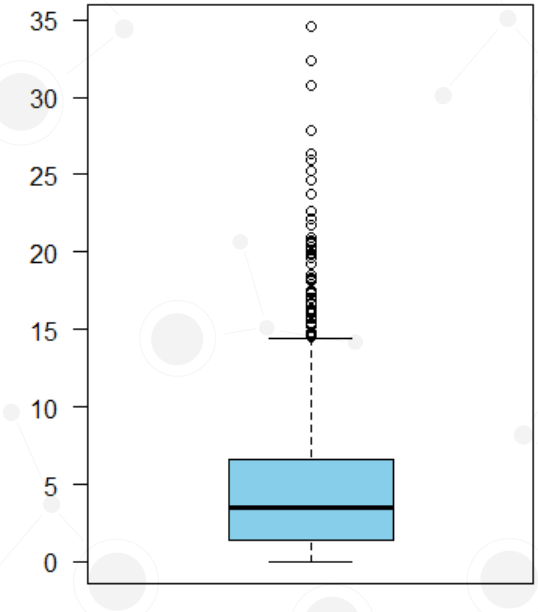
Para comprender mejor estos valores, las herramientas gráficas más usuales corresponden a:

- **Gráfico de barras discreto**, que resume la frecuencia absoluta o relativa en datos numéricos discretos.
- **Histograma**, que ilustra la frecuencia absoluta o relativa en datos numéricos continuos.
- **Diagrama de caja**, o **boxplot**, que refleja la dispersión de los datos a partir de los cuartiles.

Indicadores Gráficos

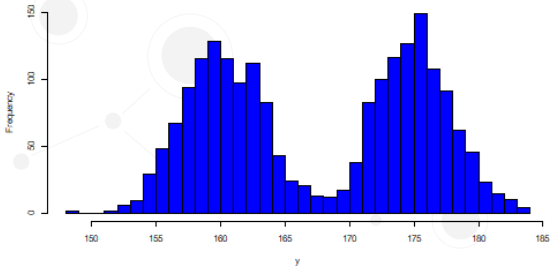
Suponga el tiempo de respuesta de un mensaje por WhatsApp

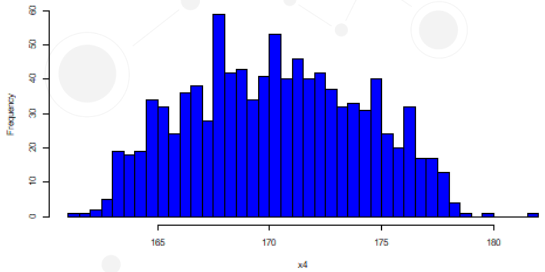


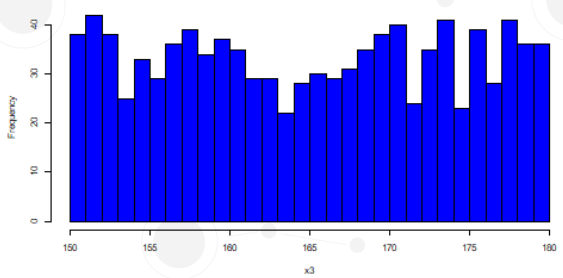


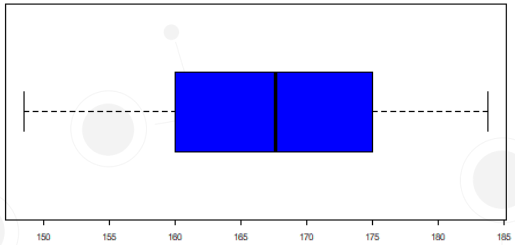
Indicadores Gráficos

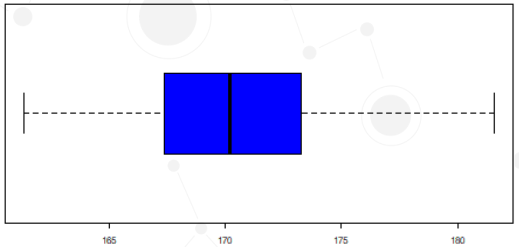
Siempre es recomendable observar ambos gráficos en conjunto, pues se puede encontrar en situaciones que no explican completamente el fenómeno.

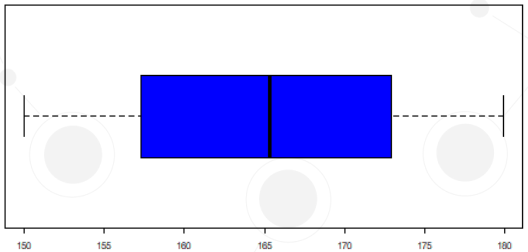






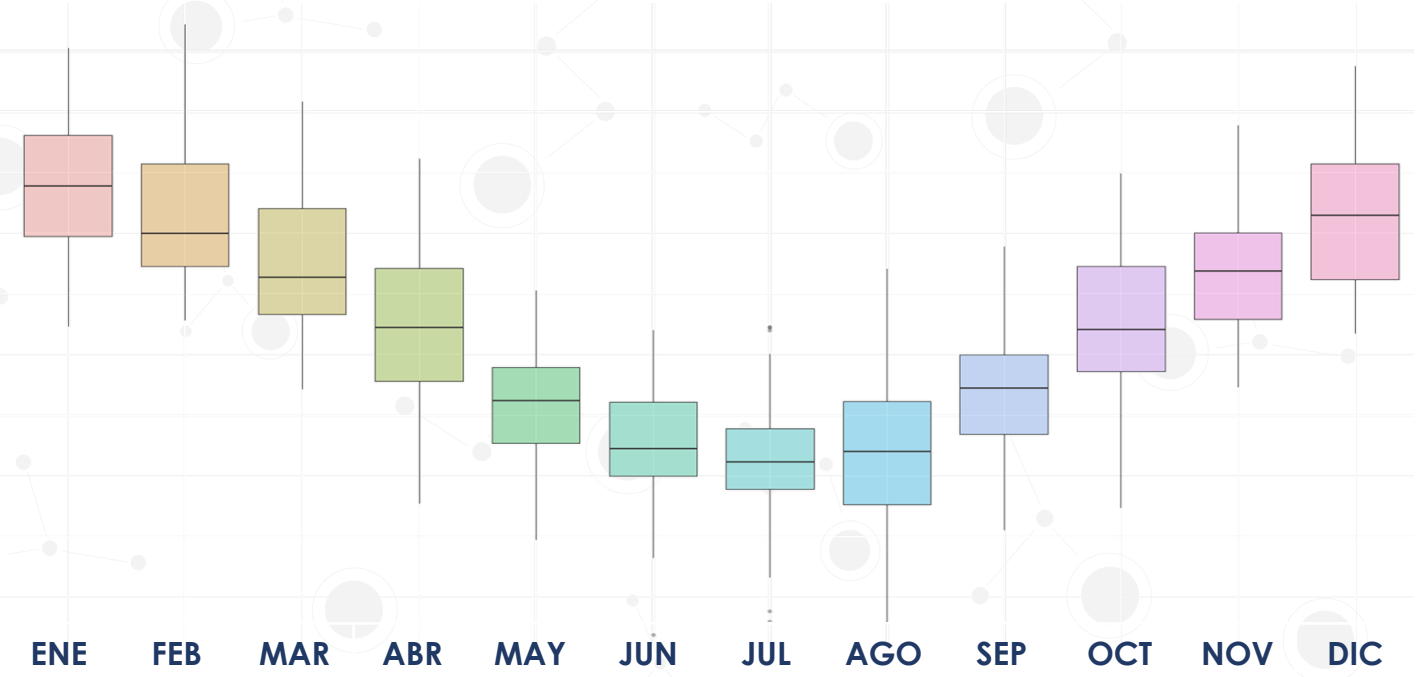






Indicadores Gráficos

La siguiente figura muestra los rangos de temperatura, para cada mes del año.



Actividad

A partir de los datos de la Encuesta nacional de salud,

I. Busque la correlación entre la TALLA y el PESO, concluya.

II. Calcule la correlación entre la EDAD y el COLESTEROL, concluya.

III. Grafique el colesterol y comente

IV. Grafique el perímetro del CUELLO, segmentado por personas que realizan deporte y las que no realizan deporte, compare.

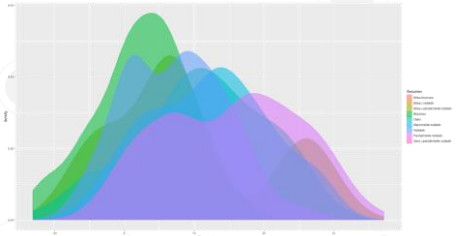
Indicadores gráficos

Los tipos de gráficos que reconoceremos son:

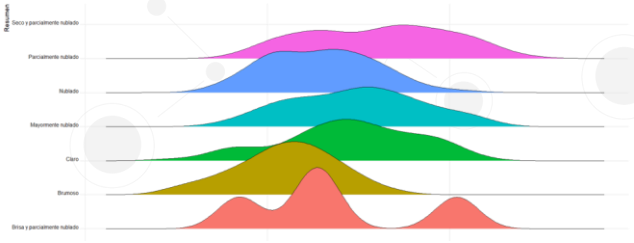
- Gráficos Distribucionales
- Gráficos de Correlación
- Gráficos de Clasificación o Ranking
- Gráficos de Participación o Parte de un todo
- Gráficos de Evolución
- Gráficos de Mapas
- Gráficos de Redes

Indicadores Gráficos

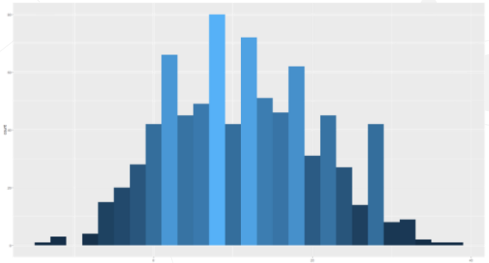
Los **gráficos distribucionales** tienen por objetivo mostrar información a partir de un comportamiento ordenado de los datos, particularmente basados en su frecuencia y similitud.



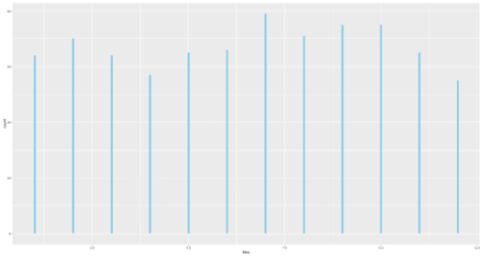
Densidad



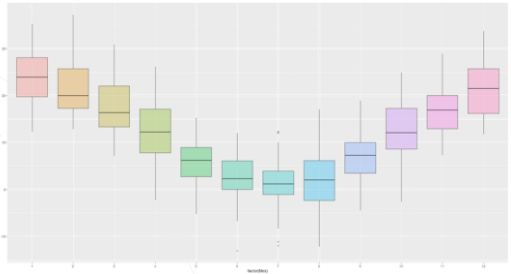
Curvas de Ridge



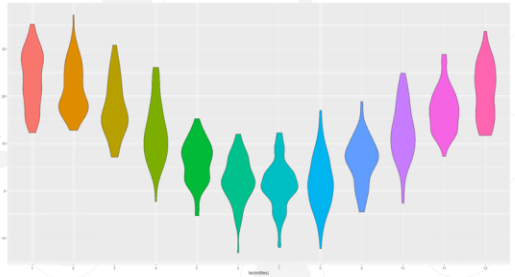
Histograma



Barras discretas



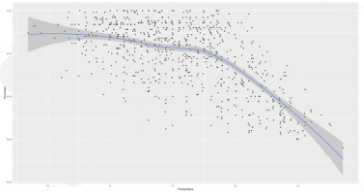
Boxplot



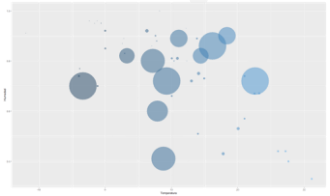
Violín

Indicadores Gráficos

Los **gráficos de correlación** están determinados para mostrar el valor de un par ordenado de valores provenientes de dos conjuntos de datos, cuya finalidad es evidenciar la relación entre dos variables.



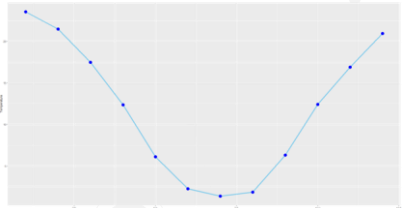
Dispersión



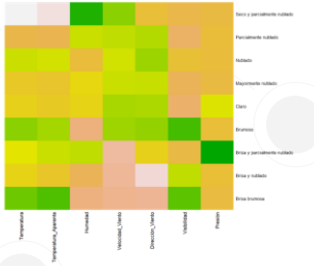
Burbuja

Correlograma

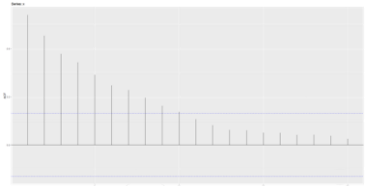
Chernov



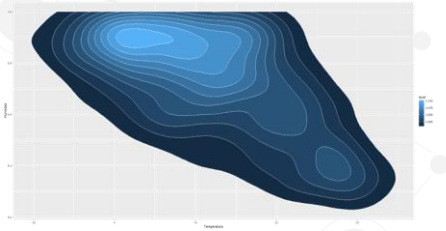
Puntos conectados



Mapas de calor



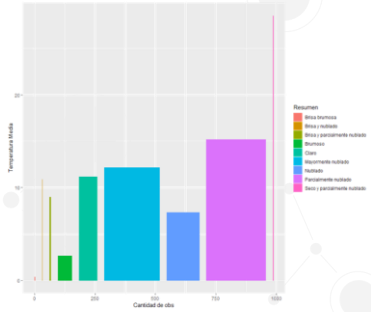
Correlograma temporal



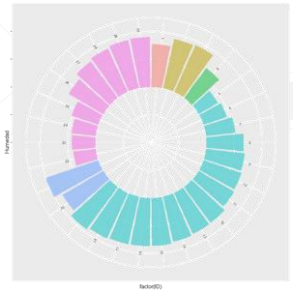
Densidad en el plano

Indicadores Gráficos

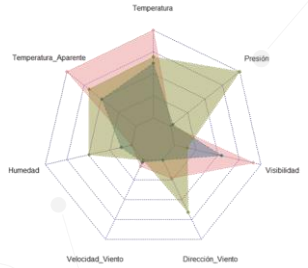
Los **gráficos de clasificación o Ranking** permiten mostrar información resumida cuando una o más variables poseen una frecuencia, principalmente en datos de categoría o ranking.



Barras

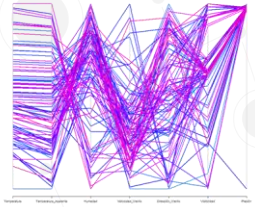


Barras circular



Radar

Nube de palabras



Líneas paralelas

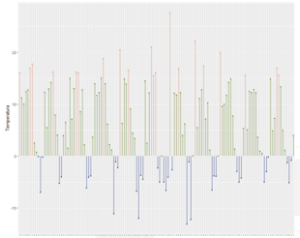


Diagrama de paletas

Indicadores Gráficos

Los **gráficos de división** o parte de un todo permiten mostrar la información de variables categóricas en cuanto a un resumen de una variable numérica, que usualmente es la frecuencia.

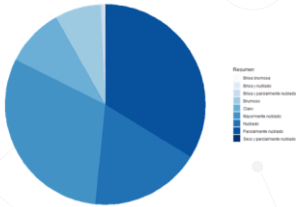
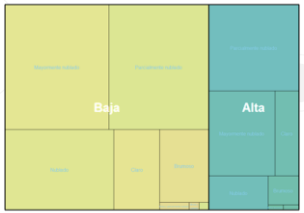
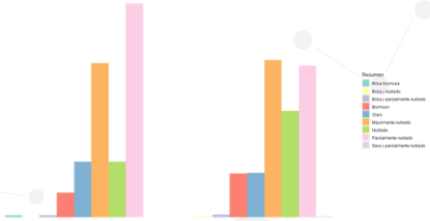


Diagrama Circular

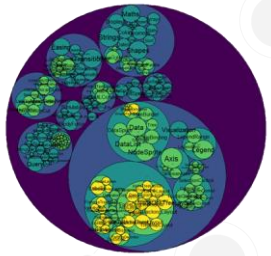


Treemap

Dendograma



Barras agrupadas

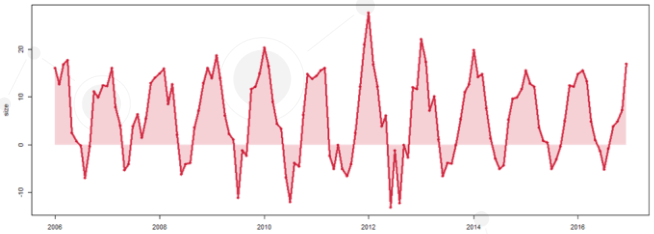


Circulares agrupados

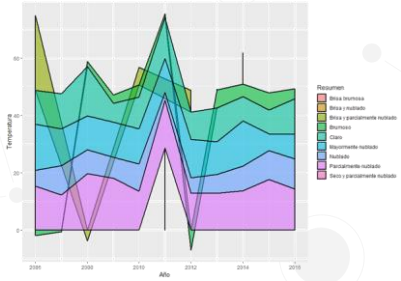
Indicadores Gráficos

Los **gráficos de evolución** están orientados a visualizar información a través de un carácter temporal u ordenado. Se clasifican según la unidad o información que se desee evidenciar.

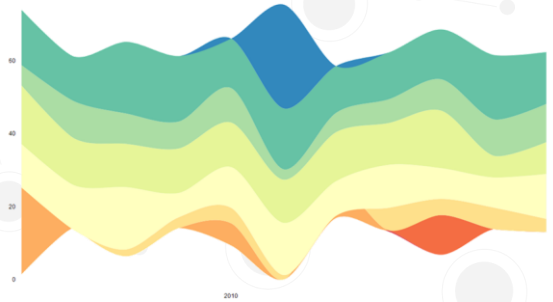
Línea temporal



Área temporal



Área agrupada

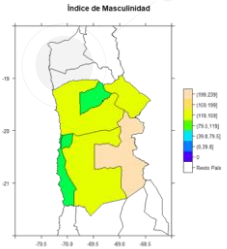


Flujo

Indicadores Gráficos

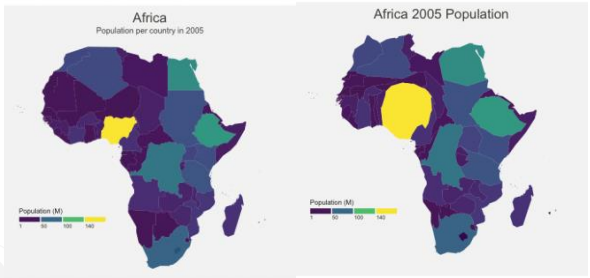
En general, se pueden utilizar mapas para mostrar diversos tipos de información. En este sección, se podrá apreciar como los **gráficos de mapas** permiten georreferenciar variables numéricas y categóricas, así como su relación, clasificación y evolución.

Posición

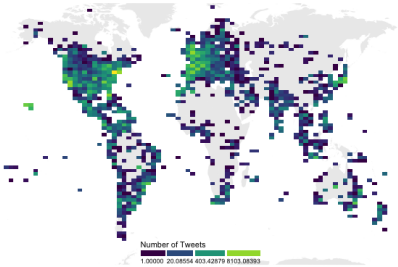


Coropleta

Mapa de Conexión



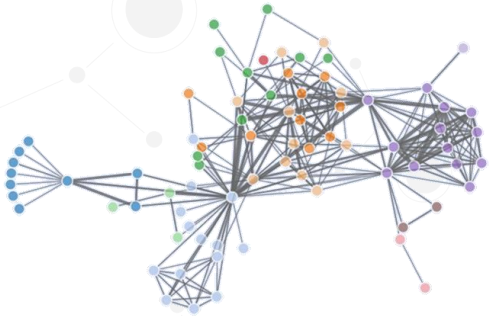
Cartograma



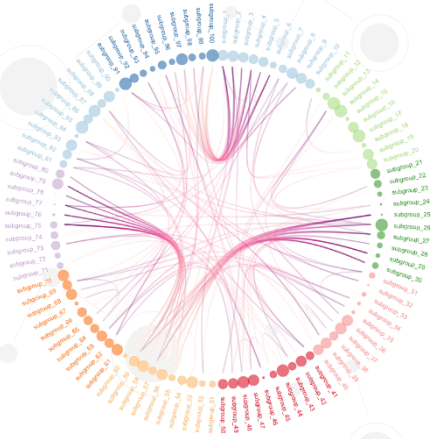
Mapa Hexagonal

Indicadores Gráficos

Los **gráficos de redes o flujo** permiten mostrar relaciones entre las categorías de una variable, a partir de otra variable numérica, que usualmente es frecuencia, correlación, medida estadística o medida contextual.



Nodos o Redes



Bordes Jerárquicos

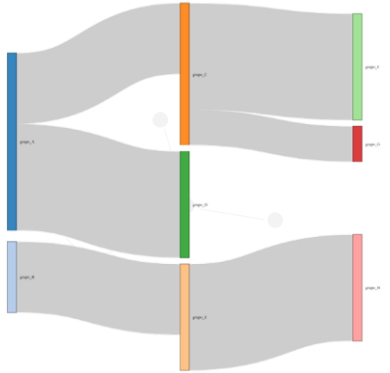
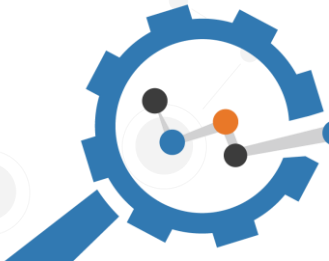


Diagrama de Sankey

Análisis de datos en EXCEL

Antes de comenzar con el entrenamiento en programación, es conveniente **manipular** herramientas con mayor alcance y más "populares", con la finalidad de tener una amplia mirada del análisis computacional, así como resultados en equipos interdisciplinarios.

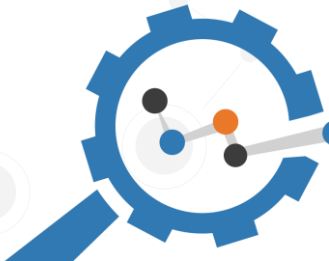
Para esto, **EXCEL** es la primera herramienta de análisis estadístico que, en casos simples, puede ser más rápido y sencillo de utilizar.



Análisis de datos en EXCEL

EXCEL es un software informático de la empresa Microsoft, que permite desarrollar actividades financieras y contables a través de hojas de cálculo.

Permite realizar manipulación de celdas con tablas, gráficos y funciones.

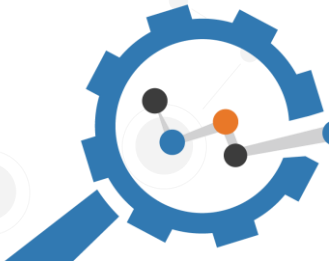


Análisis de datos en EXCEL

El método más óptimo para generar aplicaciones de análisis estadístico es a través del paquete de análisis de datos.

Para eso, se debe seguir los pasos:

- **Archivo >> Opciones >>Complementos >> Ir a**
- seleccionar la alternativa **Herramientas para el análisis**.
- Esto generará en la pestaña de datos aparecerá una nueva opción denominada **Análisis de datos**.



Actividad

Se poseen los datos de McDonald asociados a las características nutricionales del menú:

I. Explore los datos.

II. Exponga la variedad de productos por categoría.

III. Reporte un análisis numérico y gráfico del SODIO que posee el menú de McDonald.

IV. Elija dos categorías y comparé numérica y gráficamente las calorías de los productos del menú.

V. Verifique la correlación entre azúcar y proteínas.

VI. Realice un análisis exploratorio del colesterol del menú y comente sus resultados a partir del estudio de la Encuesta nacional de Salud.