

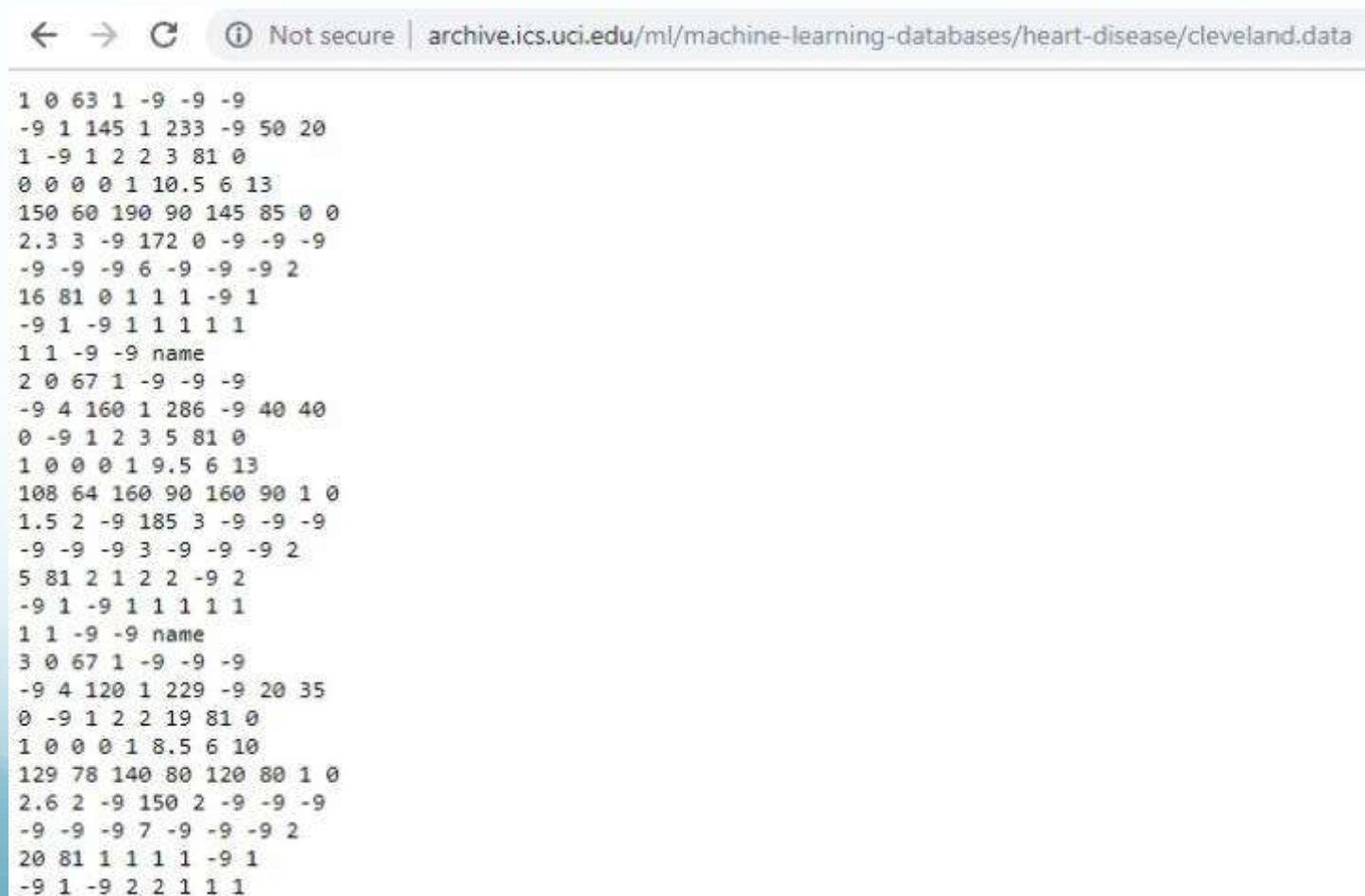
Ingeniería del Conocimiento CIF-8458

Técnicas de Visualización

Carlos Valle Vidal
Segundo Semestre 2023

¿Por qué visualizar los datos?

- Es difícil entender los datos crudos



A screenshot of a web browser displaying raw data from the Cleveland Heart Disease dataset. The browser's address bar shows the URL `archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/cleveland.data`. The page content consists of a single column of text representing the dataset, with each line containing a sequence of numerical values and some categorical labels like 'name'. The data is presented in a way that is difficult to interpret without visualization.

```
1 0 63 1 -9 -9 -9
-9 1 145 1 233 -9 50 20
1 -9 1 2 2 3 81 0
0 0 0 0 1 10.5 6 13
150 60 190 90 145 85 0 0
2.3 3 -9 172 0 -9 -9 -9
-9 -9 -9 6 -9 -9 -9 2
16 81 0 1 1 1 -9 1
-9 1 -9 1 1 1 1 1
1 1 -9 -9 name
2 0 67 1 -9 -9 -9
-9 4 160 1 286 -9 40 40
0 -9 1 2 3 5 81 0
1 0 0 0 1 9.5 6 13
108 64 160 90 160 90 1 0
1.5 2 -9 185 3 -9 -9 -9
-9 -9 -9 3 -9 -9 -9 2
5 81 2 1 2 2 -9 2
-9 1 -9 1 1 1 1 1
1 1 -9 -9 name
3 0 67 1 -9 -9 -9
-9 4 120 1 229 -9 20 35
0 -9 1 2 2 19 81 0
1 0 0 0 1 8.5 6 10
129 78 140 80 120 80 1 0
2.6 2 -9 150 2 -9 -9 -9
-9 -9 -9 7 -9 -9 -9 2
20 81 1 1 1 1 -9 1
-9 1 -9 2 2 1 1 1
```

¿Por qué visualizar los datos?

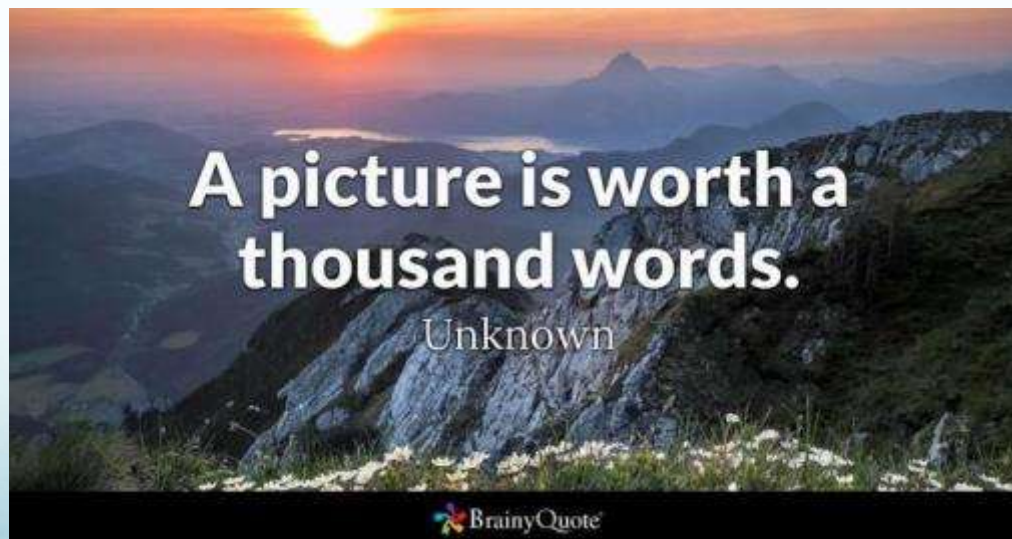
(2)

- Sin lugar a dudas, la visualización de datos es una herramienta crucial en el mundo de negocios centrado en datos de hoy.
- Algunas razones de este fenómeno son:

¿Por qué visualizar los datos?

(3)

1. La información se absorbe rápidamente: una imagen vale miles de líneas de datos. A medida que el volumen de datos aumenta inevitablemente, la visualización gestiona los flujos de nueva información y facilita la búsqueda de tendencias.



¿Por qué visualizar los datos?

(4)

2. Entender sus siguientes pasos: A partir de estas tendencias visuales, puede comprenderse más fácil sus mejores próximos pasos con menos tiempo y energía dedicados al análisis de datos.
 - Se puede ahorrar horas de tiempo mirando el panorama general en lugar de mil piezas de rompecabezas.

¿Por qué visualizar los datos?

(5)

3. Conectar los puntos: La visualización de datos no solo muestra patrones y tendencias. También pone de relieve las correlaciones y relaciones importantes pero sutiles entre las condiciones del negocio.

¿Por qué visualizar los datos?

(6)

4. Mantenga el interés de la audiencia un poco más: Mostrar gráficos de sus datos reproducen un mensaje rápidamente, antes de que pierda interés de la audiencia.
 - Como las personas ahora tienen un período de atención más corto que el de los peces dorados, mantener el interés es un objetivo crucial cuando se comparten ideas.
 - <http://time.com/3858309/attention-spans-goldfish/>

¿Por qué visualizar los datos?

(7)

5. Quite (disminuya) la necesidad de científicos de datos: La visualización de datos hace que los datos sean más accesibles y menos confusos. Hace solo unos años, los únicos profesionales que podían entender los datos de la empresa trabajaban en el departamento de TI.
 - Ahora, los equipos de finanzas, ventas y marketing no solo tienen un ávido interés en lo que los datos les dicen, sino que también tienen los medios para ir tras las respuestas.

¿Por qué visualizar los datos?

(8)

6. Permite compartir los datos: Las visualizaciones se pueden distribuir entre los equipos fácilmente, y sus equipos serán mucho más receptivos a una visual atractiva que una hoja de cálculo de Excel masiva.

¿Por qué visualizar los datos?

(9)

7. Encontrar datos atípicos: La visualización de datos revela rápidamente los valores atípicos en sus datos. Como los valores atípicos tienden a arrastrar los promedios de datos en la dirección incorrecta, es crucial encontrarlos.
 - Los gráficos rápidamente los iluminan, lo que le permite comprender **por qué están allí e ignorarlos cuando sea necesario**.

¿Por qué visualizar los datos?

(10)

8. Ayuda a recordar los puntos importantes: Los elementos visuales ayudan a enviar conceptos importantes a la memoria.
 - Es más fácil recordar y memorizar un concepto si tenemos un gráfico en el que centrarnos, no solo palabras o elementos de línea.

¿Por qué visualizar los datos?

(10)

9. Permite actuar rápidamente luego de un descubrimiento: Lo más importante es que la visualización de datos le permite descubrir cosas y tomar decisiones más rápido.
 - Usándolos, se pueden revisar las estrategias rápidamente y hacer actualizaciones de manera eficiente, ayudando a lograr el éxito con menos errores y mayor velocidad.

Herramientas para visualización en Python

- Pandas: Biblioteca que proporciona datos de alto rendimiento y fáciles de usar. Estructuras y herramientas de análisis de datos para el lenguaje de programación Python.
- Matplotlib: Biblioteca de gráficos en Python que produce figuras de calidad de publicación en una variedad de formatos impresos y entornos interactivos a través de plataformas.
- Seaborn: Biblioteca de visualización Python basada en matplotlib. Proporciona una interfaz de alto nivel para dibujo gráfico estadístico.

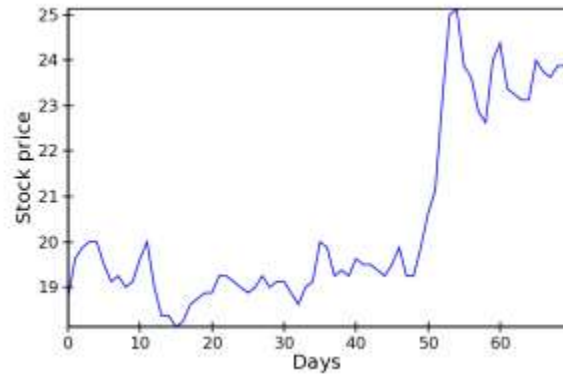
Usando dataframes con Pandas

- Como ejemplo importaremos el dataset Pima indians Diabetes.
- Contiene información de 768 pacientes.
- Los atributos de cada paciente son:
 - Pregnancies: Número de embarazos
 - GlucosePlasma: concentración de glucosa a 2 horas de un exámen oral de tolerancia a la glucosa.
 - BloodPressure: Presión sanguínea diastólica (mm Hg)
 - SkinThicknessTriceps: skin fold thickness (mm)
 - Insulin2-Hour: serum insulin (mu U/ml)
 - BMIBody mass index (weight in kg/(height in m)^2)
 - DiabetesPedigreeFunction: Diabetes pedigree function
 - Age: Edad (años)
 - OutcomeClass: Tiene o no diabetes (0 or 1)

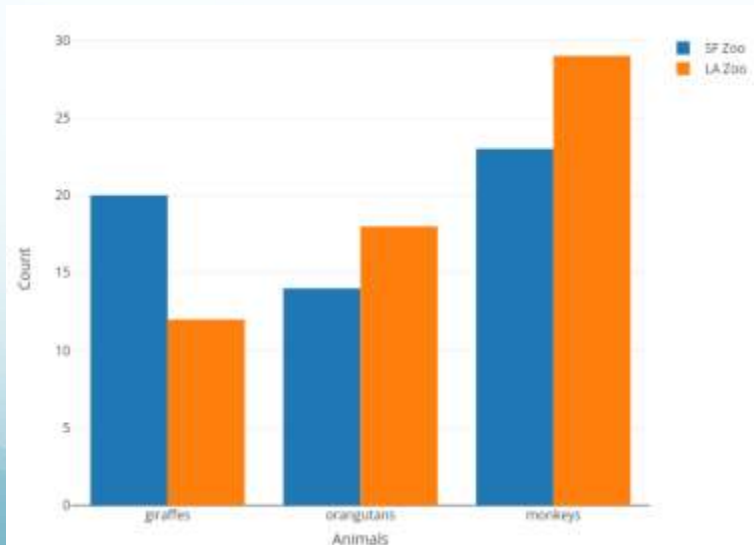
Usando dataframes con Pandas (2)

- Ver Jupiter notebook

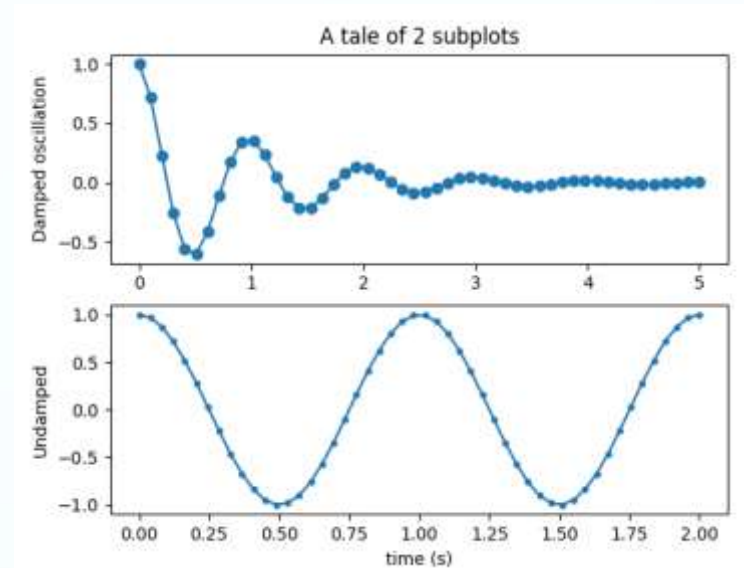
Tipos de gráficos



Line plot

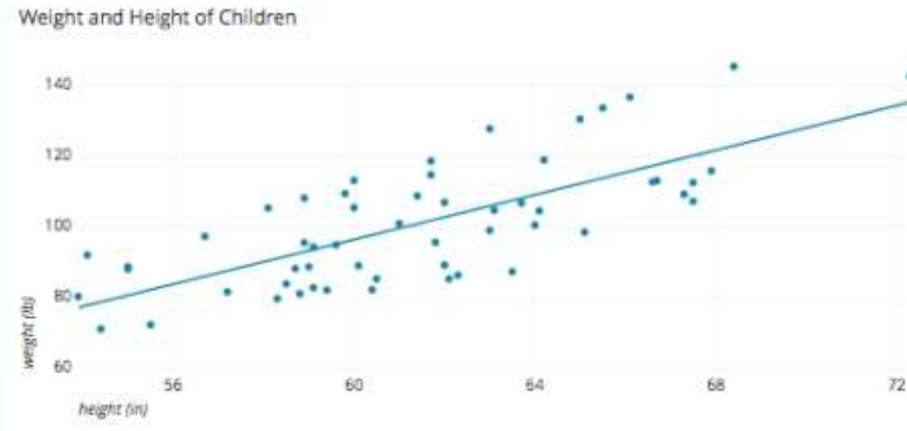


Bar charts

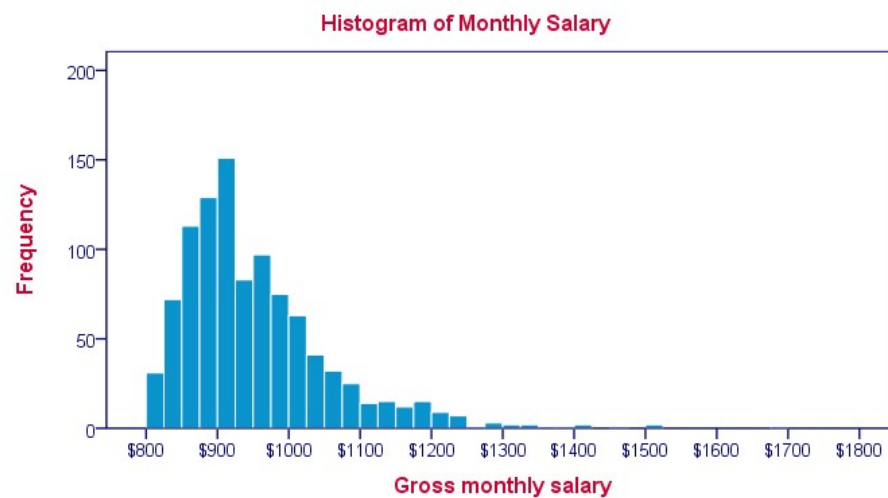


Subplots

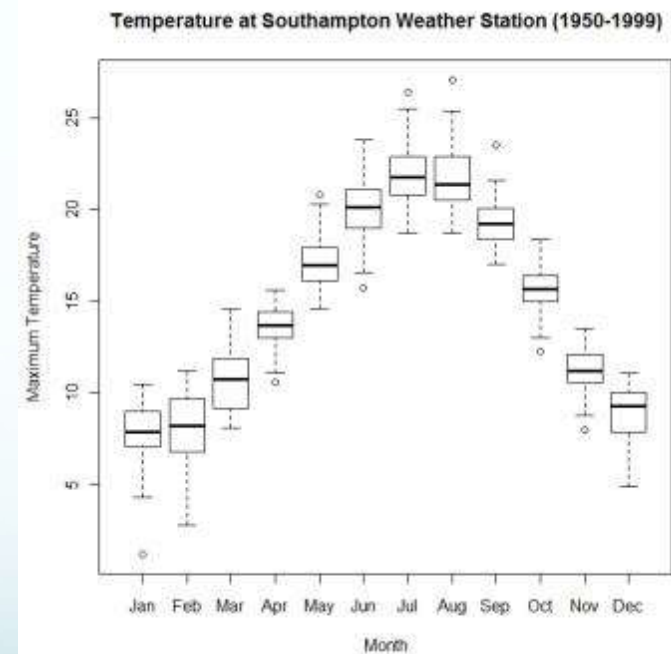
Tipos de gráficos (2)



Scatter plot



Histograms



Box and whiskers plot

Gráfico de líneas

- Line Plot o gráfico de líneas: Los gráficos de líneas son perfectos para mostrar tendencias a lo largo de un período de tiempo.

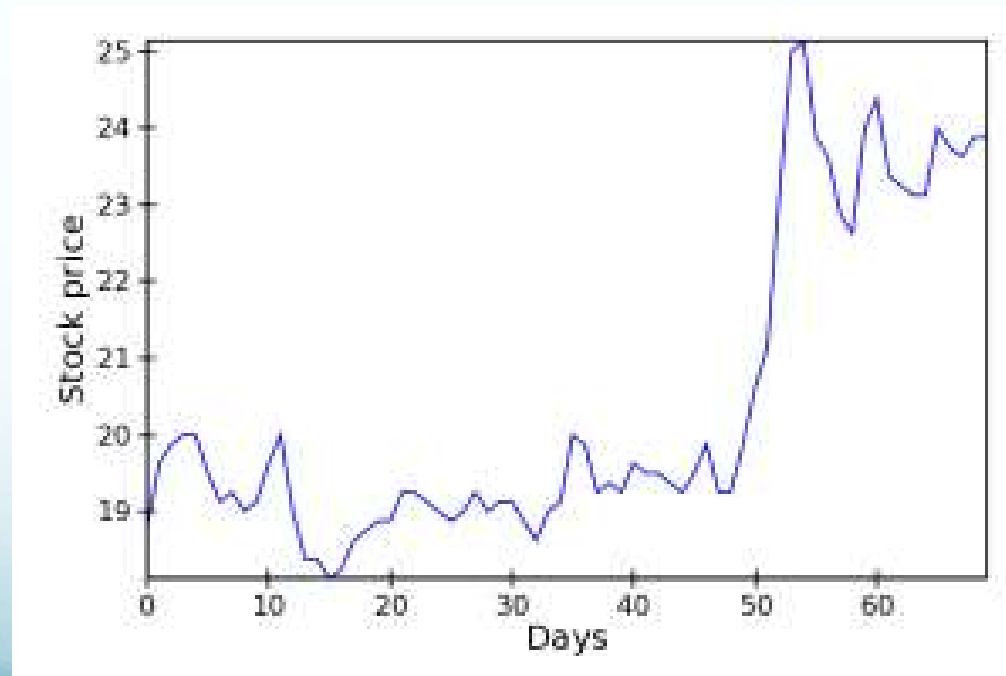


Gráfico de barras

- Bar Chart o grafico de barras: Los gráficos de barras permite una mejor comprensión del gráfico apilando la información en barras. El gráfico de barras también es útil cuando queremos comparar valores uno al lado de otro y cuando queremos visualizar variables que se calculan empleando la misma unidad.
- El gráfico de barras no funciona tan bien cuando hay muchos valores de dimensión debido a las limitaciones en la longitud de los ejes.

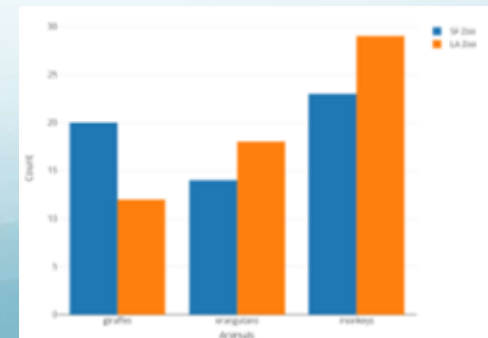


Gráfico de dispersión

- Scatter plots o gráfico de dispersión: Los gráficos de dispersión sirven para identificar correlaciones (o dependencias) entre las variables. Es muy útil para contrastar las correlaciones de distintos grupos de datos. Por ejemplo: contrastar la relación venta-precio de ξ

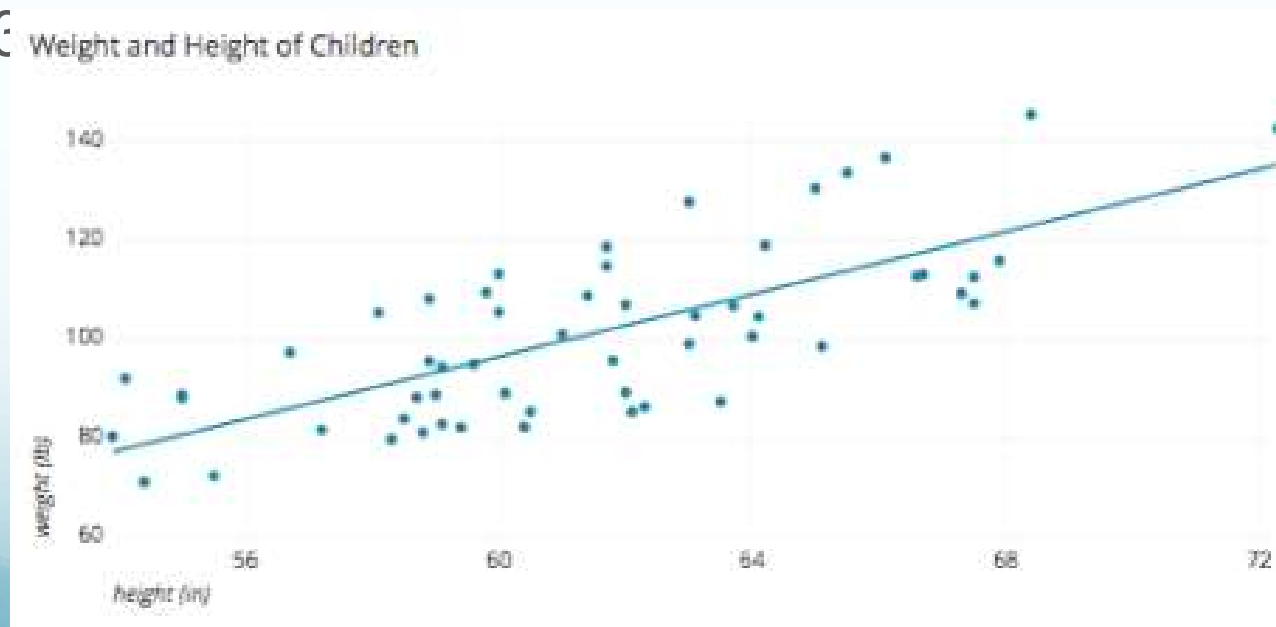
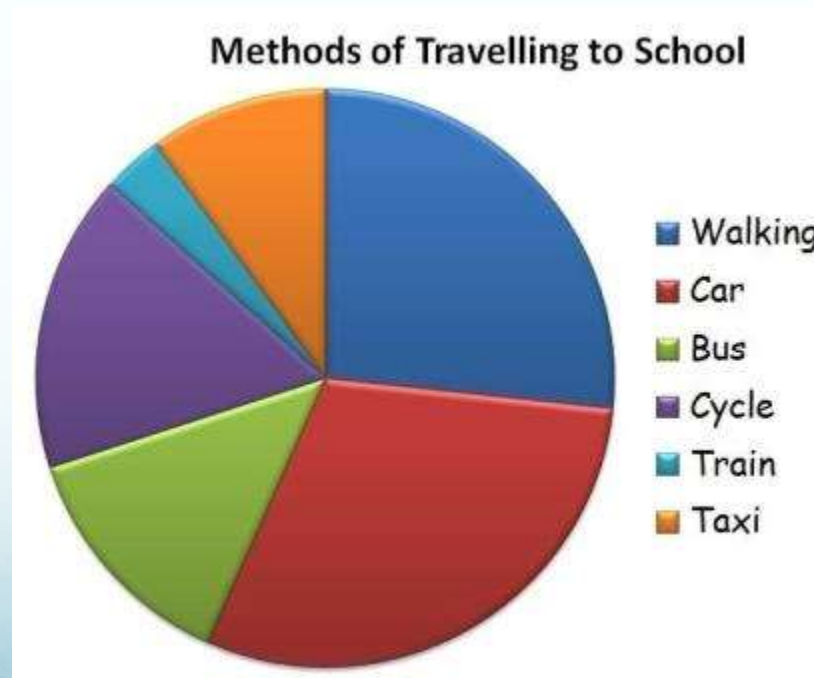


Gráfico de tortas

- Pie Chart o gráfico de tortas: Muy útil para comparar proporciones de valores que se repiten de una determinada característica.



Histogramas

- En estadística, un histograma representa gráficamente los valores que ha tomado una variable usando barras. Donde la superficie de cada barra es proporcional a la frecuencia de los valores representados. En el eje vertical se muestran las frecuencias y en el eje horizontal los valores de las variables.

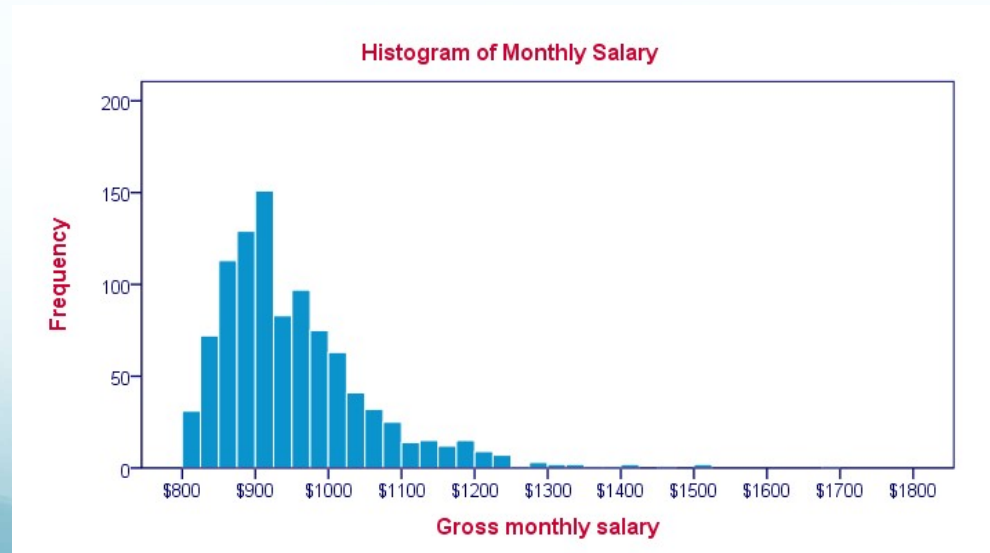


Grafico de densidad

- Density Plots o *gráfico de densidad de Kernel*: Un gráfico de densidad visualiza la distribución de datos en un intervalo o período de tiempo continuo. Este gráfico es una variación de un histograma usando un método de suavizado para trazar valores.
- Los picos de un gráfico de densidad ayudan a mostrar dónde los valores se concentran en el intervalo.

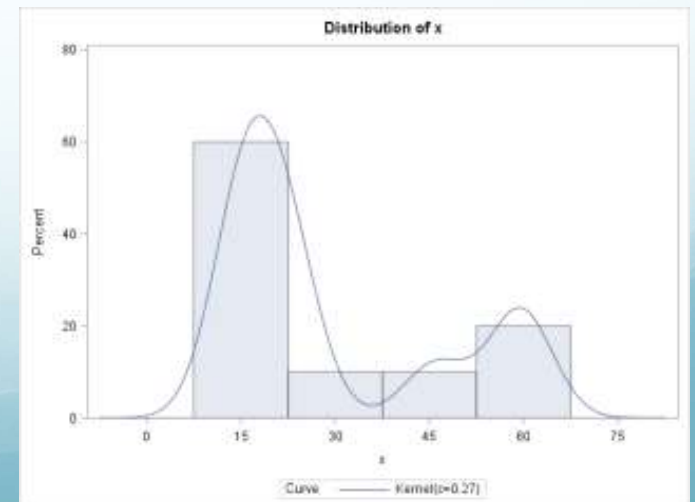


Diagrama de cajas y bigotes

- Box and Whisker Plots o box plot: Es un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles.
- De esta manera, el diagrama de caja muestra a simple vista la mediana y los cuartiles de los datos¹, pudiendo también representar los valores atípicos de estos.

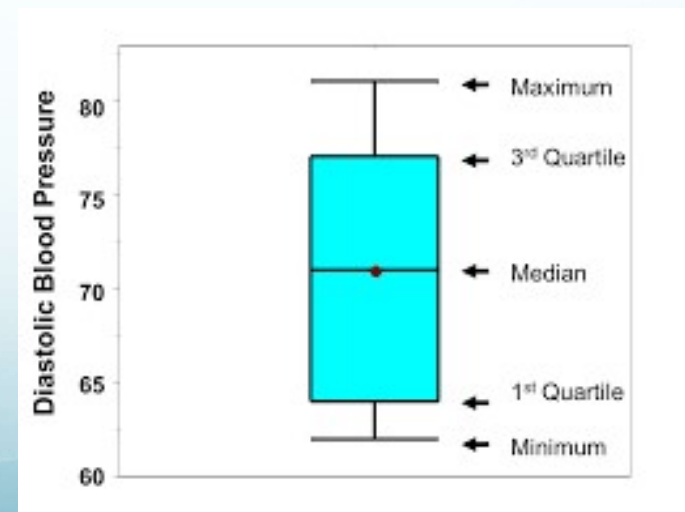
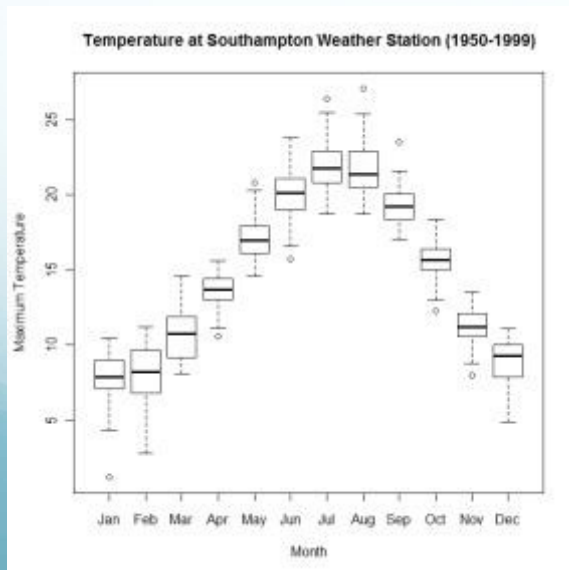
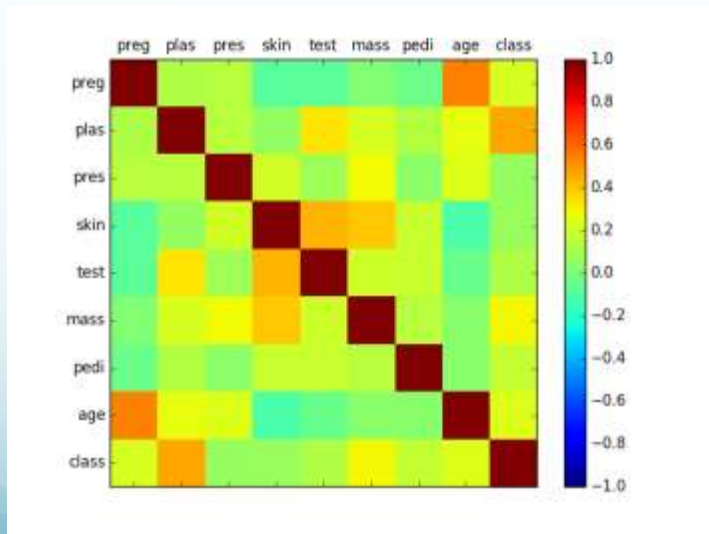
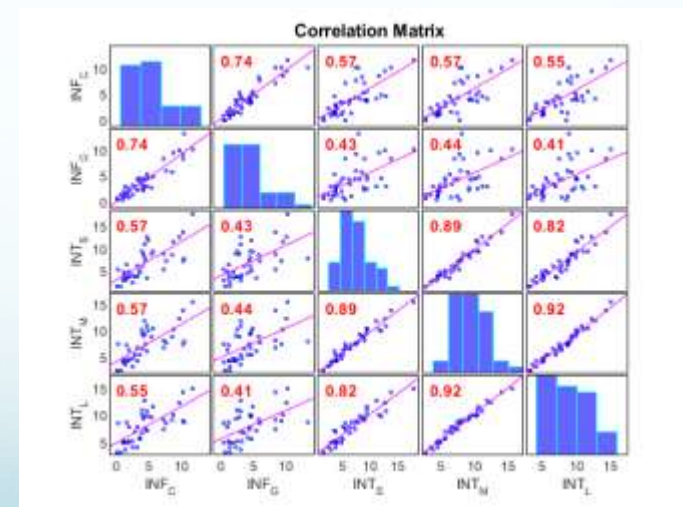


Diagrama de correlación matricial (calor)

- Correlation Matrix Plot: La matriz de correlación, se utiliza para investigar la dependencia lineal entre múltiples variables al mismo tiempo. El resultado es una tabla que contiene los coeficientes de correlación entre cada variable y las otras.



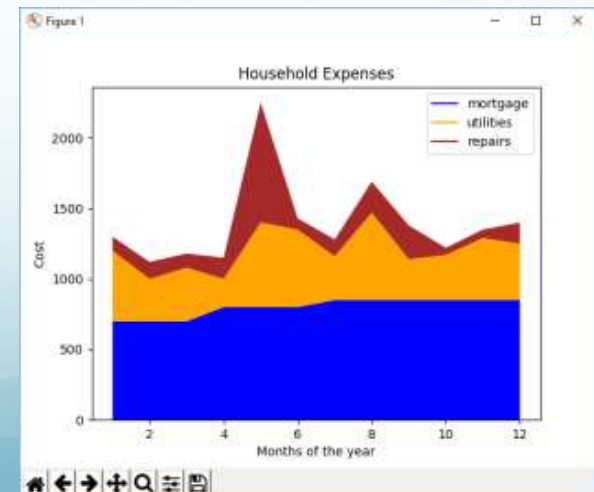
Heat map



Scatter plot

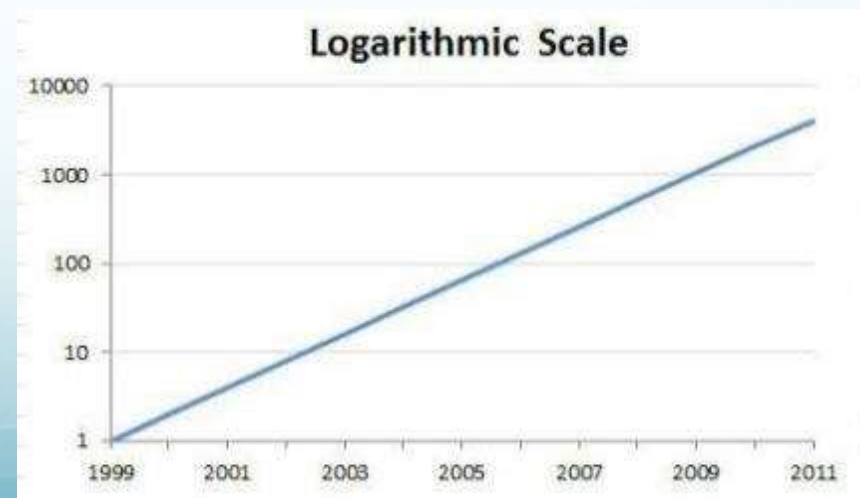
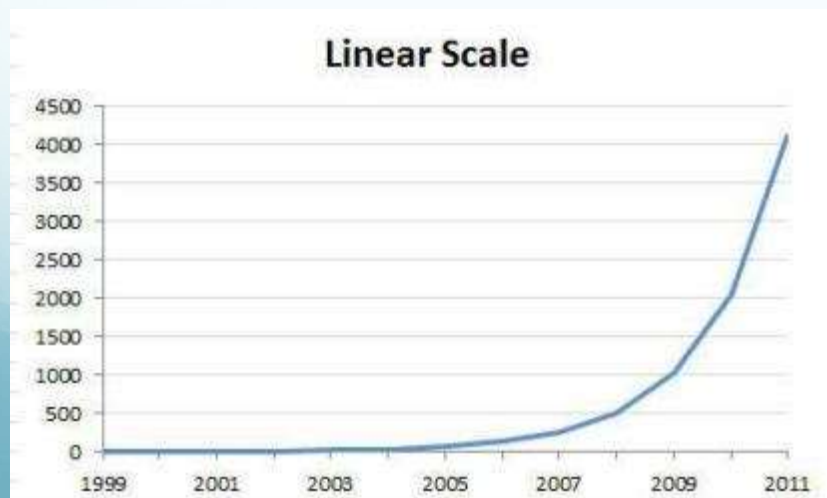
Stack plots

- Stack plots: Un gráfico de pila o stack plot es un gráfico que muestra todo el conjunto de datos con una fácil visualización de cómo cada parte conforma el conjunto.
- Cada constituyente del gráfico se apila uno encima del otro.
- Muestra la composición parcial de la unidad, así como toda la unidad.



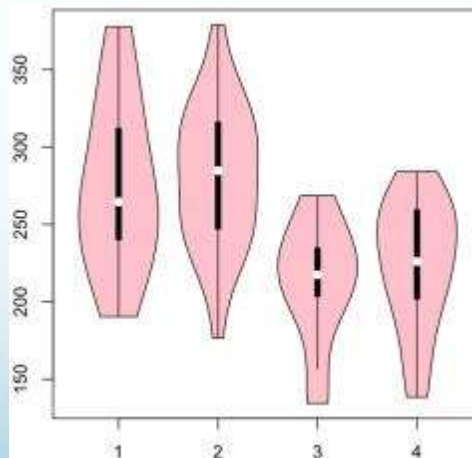
Log plots

- Los gráficos en escala logarítmica son útiles cuando existe asimetría hacia valores grandes; es decir, casos en los que uno o unos pocos puntos son mucho más grandes que la mayor parte de los datos.
- También se usan para mostrar el cambio porcentual o factores multiplicativos.



Violin plots

- Violin plots: El diagrama de violín es un método mostrar datos que mezcla el boxplot con el diagrama de densidad, mostrando la densidad de probabilidad de los datos a diferentes valores de (en el caso más simple esto podría ser un histograma).



Tipos de gráficos

- Mas detalles:
 - Jupiter notebook.