



# Tema 2:

## Técnicas de Visualización – Parte I/2

**2.1 Introducción**

**2.2 Datos de 2 y 3 dimensiones**

**2.3 Técnicas de Visualización de datos multivariados**

2.3.1 Proyección Geométrica

Máster Universitario en Ingeniería Informática

Pablo.Bermejo@uclm.es

## 2.1 Introducción

- **Visualización de la Información** es el uso de representaciones visuales de un conjunto de datos en un formato de 2 dimensiones (papel, pantalla), con el objetivo de amplificar nuestro conocimiento acerca del mismo:
  - Descubrir conocimiento inesperado
  - Profundizar en lo esperado
- Normalmente un conjunto de datos es un grupo de registros descritos cada uno por  **$n$  variables**:
  - $n \leq 3$  representación directa
  - $n > 3$  técnicas de visualización multivariada; su comprensión no es directa y requieren entrenamiento.

## 2.1 Introducción

- Utilizar técnicas efectivas de visualización de datos multivariados nos permitirá *identificar, distinguir, categorizar y asociar* las relaciones subyacentes entre las variables o registros.
- Encontrar el mejor método de visualización para un conjunto de datos concreto es un problema **difícil y no determinista**. Si los datos son multivariados, además nos encontramos con los siguientes **problemas**:
  - Proyección: la técnica utilizada para mapear los datos a 2 dimensiones puede **saturar** la capacidad de observación del destinatario. Es importante que las variables puedan observarse en **conjunto**, mientras que al mismo tiempo puedan juzgarse de forma **independiente**.
  - Dimensionalidad: cuantas más variables presenten nuestros registros, menos efectivas serán las técnicas de visualización; de ahí la importancia de la **Selección** de Variables. Además, distinto **orden** en las mismas variables puede resultar en diferencias fundamentales en la percepción visual.

## 2.1 Introducción

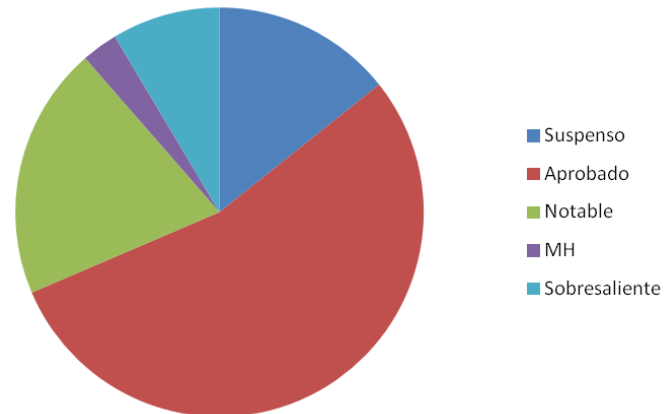
- Balance: de los anteriores problemas, se desprende que es necesario dedicar un tiempo a decidir cuánta cantidad de **información** queremos mostrar, en perjuicio de la **simplicidad** del modelo.
- Evaluación: el principal objetivo de la visualización es encontrar patrones o agrupamientos desconocidos. Entonces, puesto que no sabemos qué esperar encontrar, no existen métodos objetivos para evaluar cómo de efectivo es una técnica. Lo más directo es comparar técnicas entre sí.
- **Multidimensional Vs. Multivariado**:
  - El término *multivariado* hace referencia a un registro de alta dimensionalidad
  - El término *multidimensional* hace referencia a una variable discreta con más de 2 valores posibles.

## 2.2 Datos de 2 y 3 dimensiones

- **Diagrama de sectores, pastel o circular:**

- Representar porcentajes correspondientes a la **frecuencia** de cada posible valor de una **variable categórica**.
- El tamaño de cada sector representa el porcentaje, siendo 100% el total
- Efectivo para comparar proporciones dentro de un mismo diagrama de sectores
- Inadecuado para comparar entre distintos diagramas de sectores

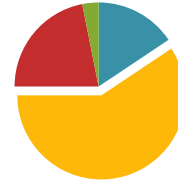
Nota	Cantidad
Suspenso	5
Aprobado	19
Notable	7
Sobresaliente	3
MH	1



## 2.2 Datos de 2 y 3 dimensiones

- Seccionado o explotado:

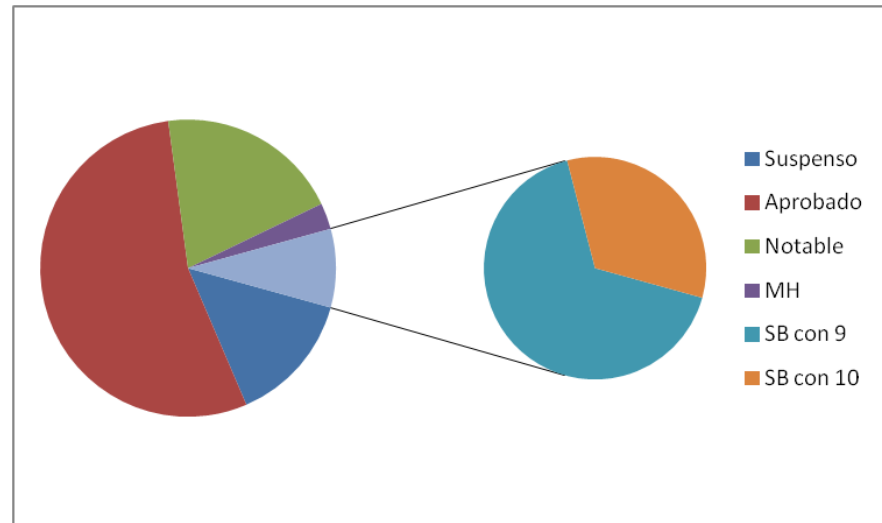
- Separa 1 ó más sectores
- Resaltar una proporción



- Con subdiagrama circular:

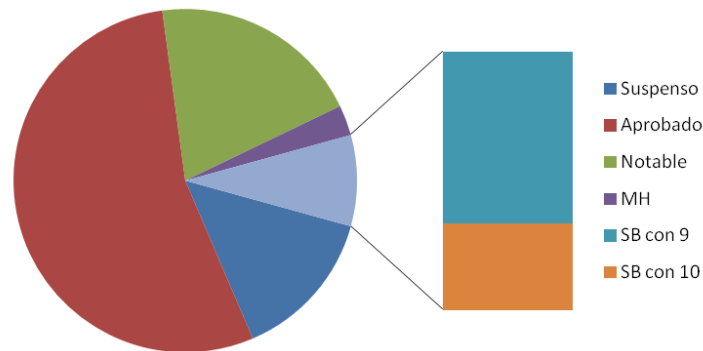
- Permite desglosar una variable en varias

Nota	Cantidad
Suspenso	5
Aprobado	19
Notable	7
MH	1
SB con 9	2
SB con 10	1

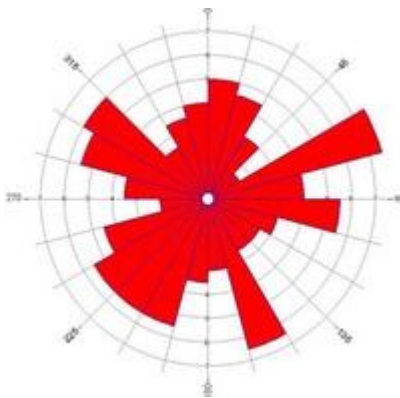


## 2.2 Datos de 2 y 3 dimensiones

- Con subdiagrama de barras:
  - Igual que el anterior, siendo el subgráfico una diagrama de barras que también representa porcentaie.



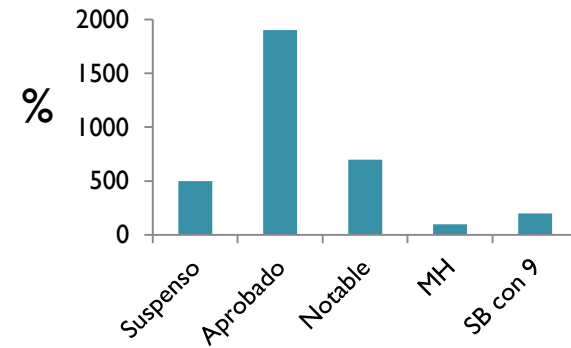
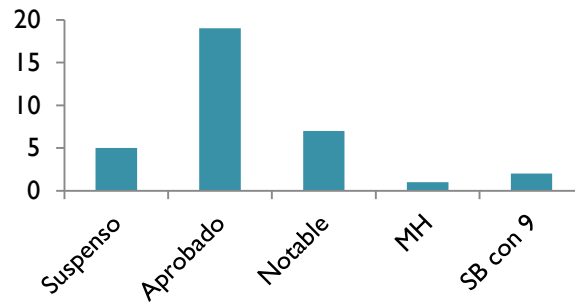
- De área polar:



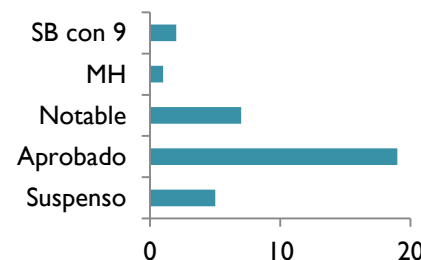
- Mismo ángulo
- La cantidad se representa por la altura

## 2.2 Datos de 2 y 3 dimensiones

- **Diagrama de barras o columnas:**
- Simple:
  - El eje horizontal representa los posibles valores de una **variable categórica**
  - El eje vertical representa la **cantidad de veces o frecuencia** que aparece de cada valor en nuestra base de datos.



- Las barras pueden ser horizontales, en cuyo caso los papeles de los ejes horizontal y vertical se intercambian.

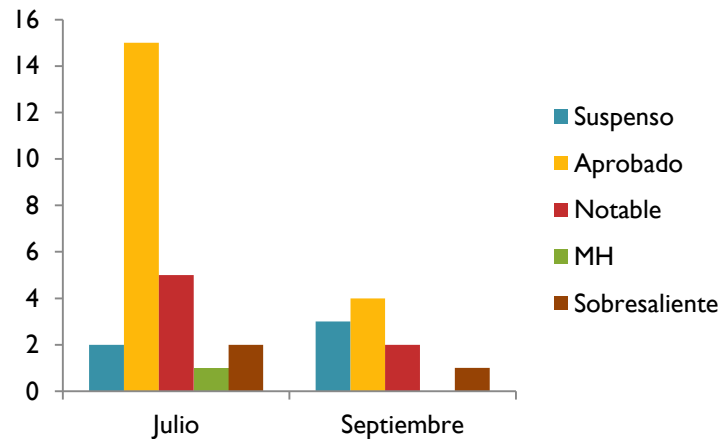




## 2.2 Datos de 2 y 3 dimensiones

- Agrupadas:
  - Contrastar una variable categórica en función de otra variable categórica

	Suspenso	Aprobado	Notable	MH	Sobresaliente
Julio	2	15	5	1	2
Septiembre	3	4	2	0	1

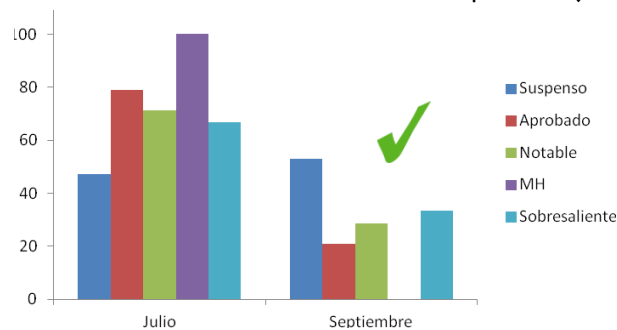


## 2.2 Datos de 2 y 3 dimensiones

- Si se utilizan **frecuencias** y se quiere comparar entre valores de la variable de referencia (Fecha en el ejemplo), hay que calcularlas de forma relativa en vez de absoluta dentro de un subgrupo, ya que el objetivo es comparar entre categorías.

	Suspense	Aprobado	Notable	MH	Sobresaliente
Julio	8	15	5	1	2
Septiembre	9	4	2	0	1

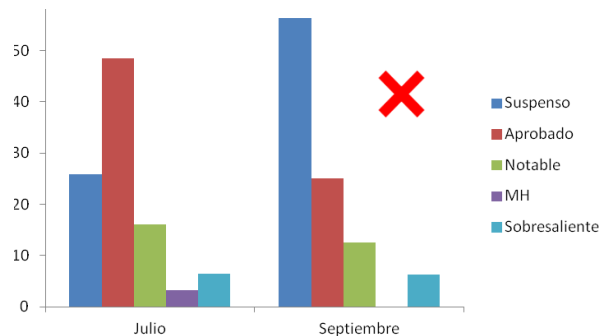
- Relativa:** Suspense en julio  $(8 \cdot 100) / (8 + 9) = 47.05$
- Absoluta:** Suspense en julio  $(8 \cdot 100) / (8 + 15 + 5 + 1 + 2) = 25.80$



Julio  
Septiembre

Suspense	Aprobado	Notable	MH	Sobresaliente
47.0588235	78.9473684	71.4285714	100	66.6666667
52.9411765	21.0526316	28.5714286	0	33.3333333

*‘Los suspenses del curso se reparten igual en las 2 convocatorias’*



Julio  
Septiembre

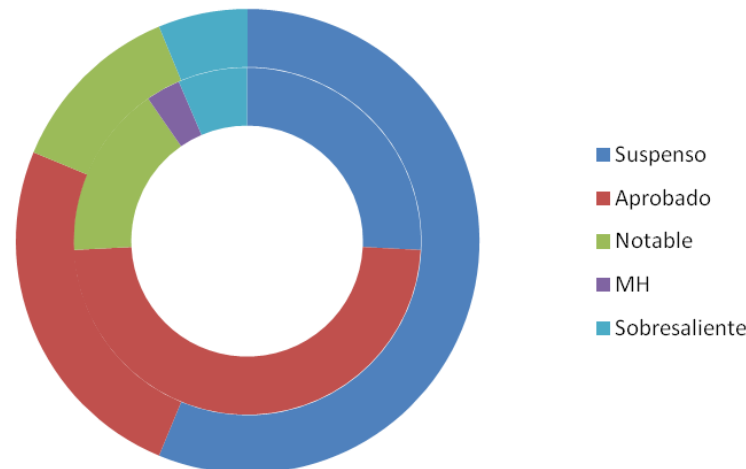
Suspense	Aprobado	Notable	MH	Sobresaliente
25.8064516	48.3870968	16.1290323	3.22580645	6.4516129
56.25	25	12.5	0	6.25

## 2.2 Datos de 2 y 3 dimensiones

- **Diagrama de anillo**

- Como el diagrama de sectores, pero añade una segunda variable categórica como referencia: notas en Julio y Septiembre.
- ¡Ojo!, las frecuencias de cada anillo son absolutas al valor correspondiente de la variable categórica

	Suspense	Aprobado	Notable	MH	Sobresaliente
Julio	8	15	5	1	2
Septiembre	9	4	2	0	1

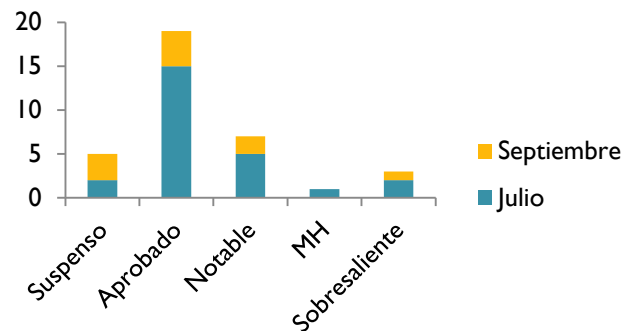


*‘Han suspendido en mucha mayor proporción los presentados en Septiembre que en Julio’*

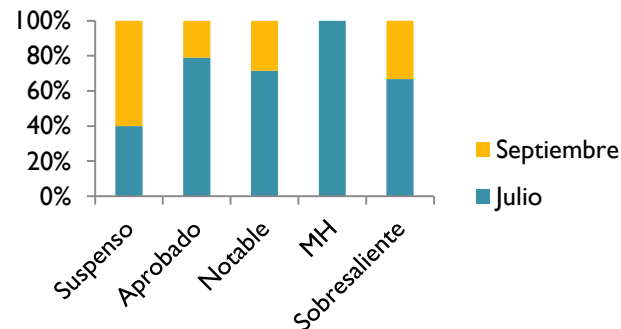
## 2.2 Datos de 2 y 3 dimensiones

- Apiladas o compuestas

- El mismo objetivo que las agrupadas, pero en vez de hacer grupo de barras por cada valor de la variable de referencia, estas se apilan en cada valor de las variable secundaria.



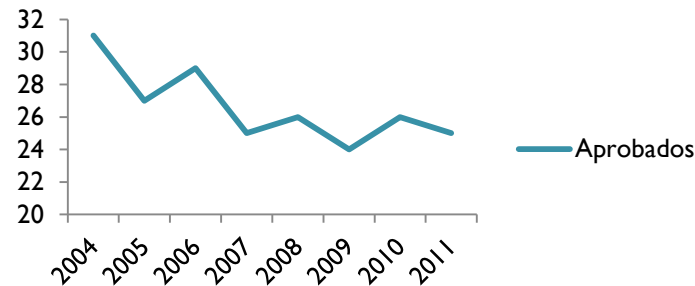
- También se pueden apilar las frecuencias relativas:



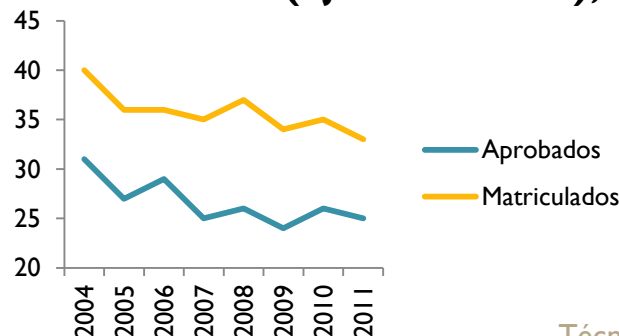
## 2.2 Datos de 2 y 3 dimensiones

- **Diagrama de líneas**

- Valor de una variable numérica según cambia el valor de otra variable numérica
- Muy útil:
  - Para observar tendencias a lo largo del tiempo en intervalos iguales
  - Cuando el orden es importante



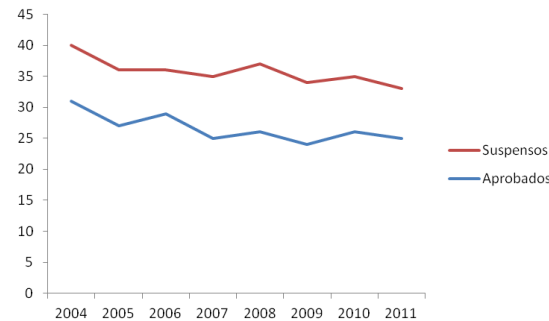
- Se puede observar la evolución de varias **variables (eje vertical) respecto a la variable de referencia (eje horizontal)**, todas ellas siendo numéricas.



## 2.2 Datos de 2 y 3 dimensiones

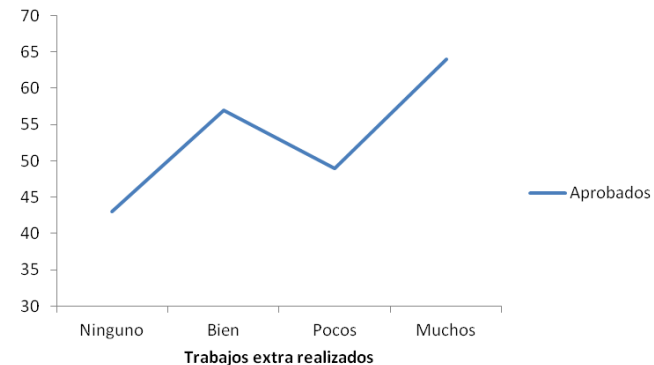
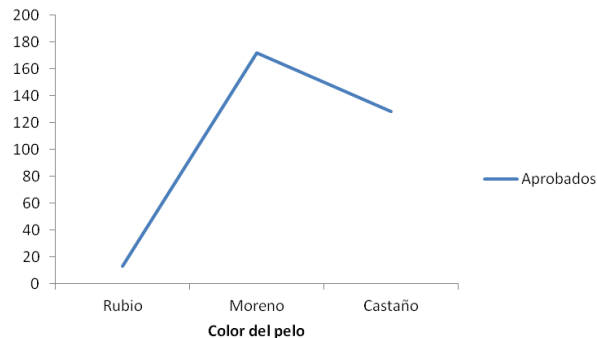
- Apiladas

- Muestra la cantidad con que aporta cada variable secundaria en cada punto de la variable de referencia
- Las barras apiladas son más sencillas de leer y representan lo mismo



- Variable de referencia categórica:

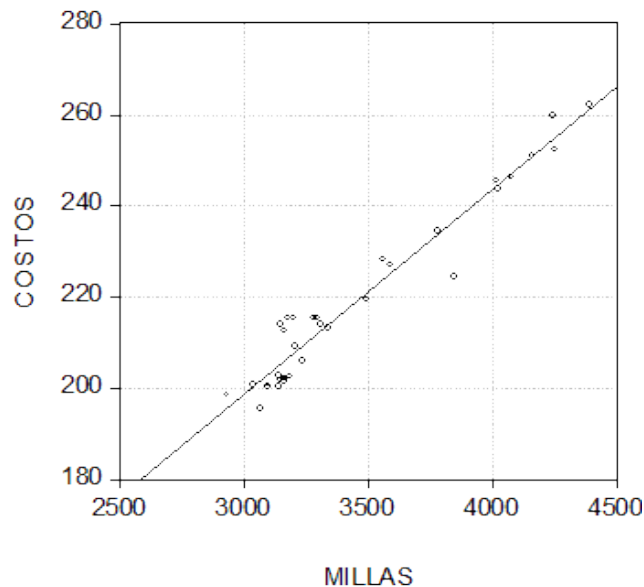
- Ya no se muestra tendencia pero sí patrones (al no ser que la variable categórica tenga carácter ordinal)



## 2.2 Datos de 2 y 3 dimensiones

- **Diagrama de dispersión**

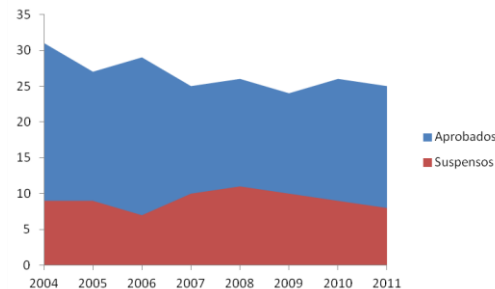
- Son más útiles que las líneas cuando hay muchos puntos próximos y queremos marcarlos en vez de unirlos.
- Además de tendencias, pueden mostrar patrones o clusters.
- Si se imprimen sobre una recta de regresión, nos indican el grado de error cometido por la ecuación.



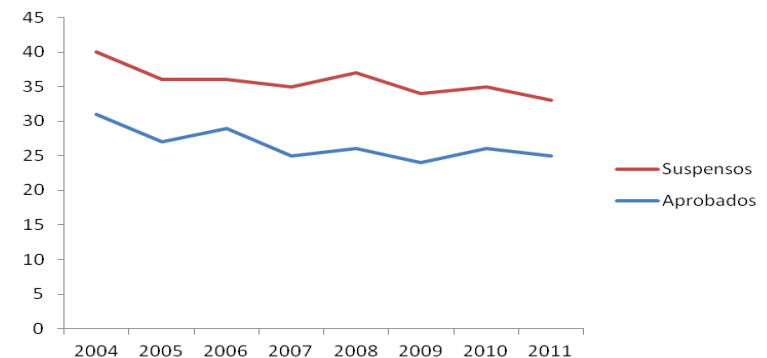
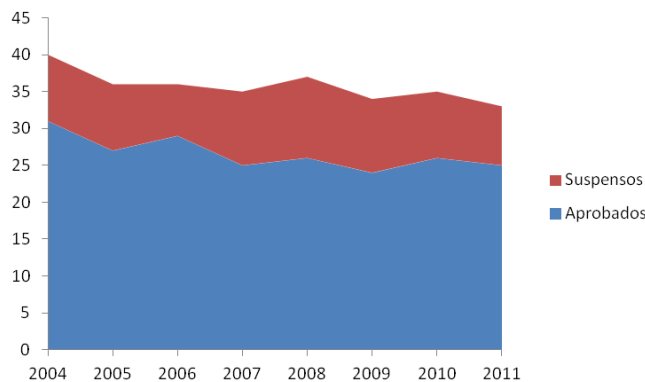
## 2.2 Datos de 2 y 3 dimensiones

- **Diagrama de áreas**

- Misma función que el de líneas, pero **puede dar la falsa impresión de que las variables están apiladas**. No son recomendables.



- Áreas apiladas



- También se pueden apilar frecuencias, como en las anteriores.

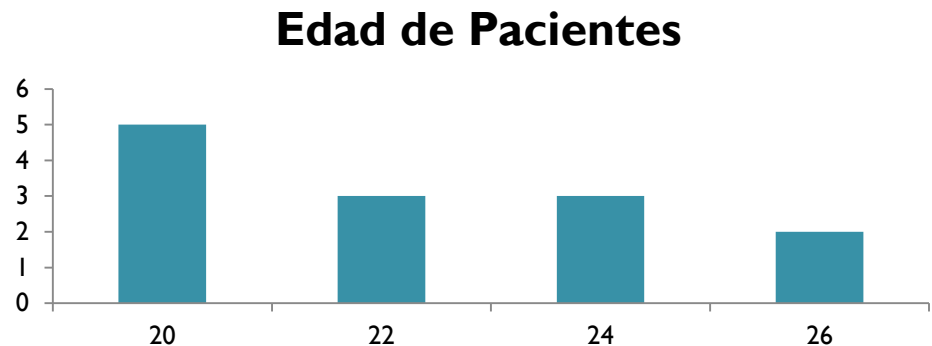


## 2.2 Datos de 2 y 3 dimensiones

- **Histogramas**

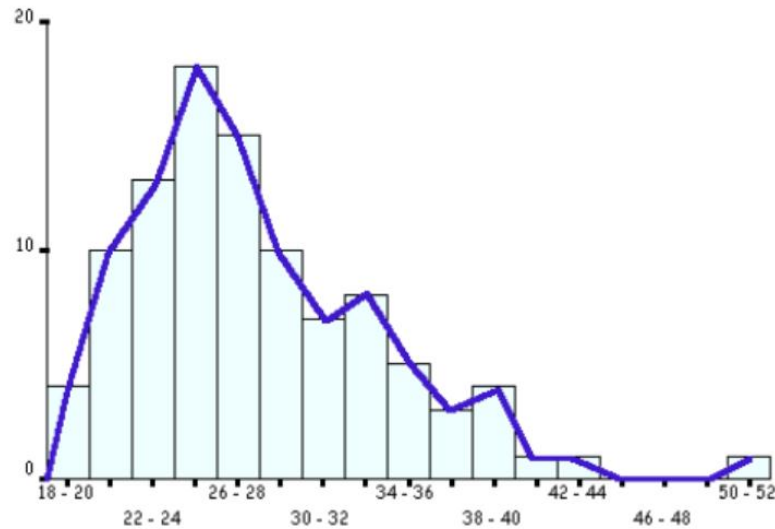
- Representación con barras verticales de **una variable continua**
- Los valores de la variable se dividen en igual amplitud
- Para cada rango, una columna vertical con altura que indica **frecuencia** o **cantidad de veces** en las que ese rango de valores se da en la base de datos.

Edad	Rango Valores
18	20
18.5	22
19	24
20	26
20.6	
20	
21	
21.6	
23.5	
23	
23.7	
24.2	
24.5	



## 2.2 Datos de 2 y 3 dimensiones

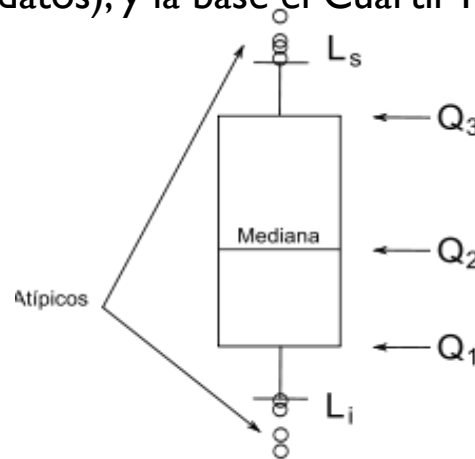
- Normalmente las barras se dibujan agrupadas, y se traza el **polígono de frecuencias**.



## 2.2 Datos de 2 y 3 dimensiones

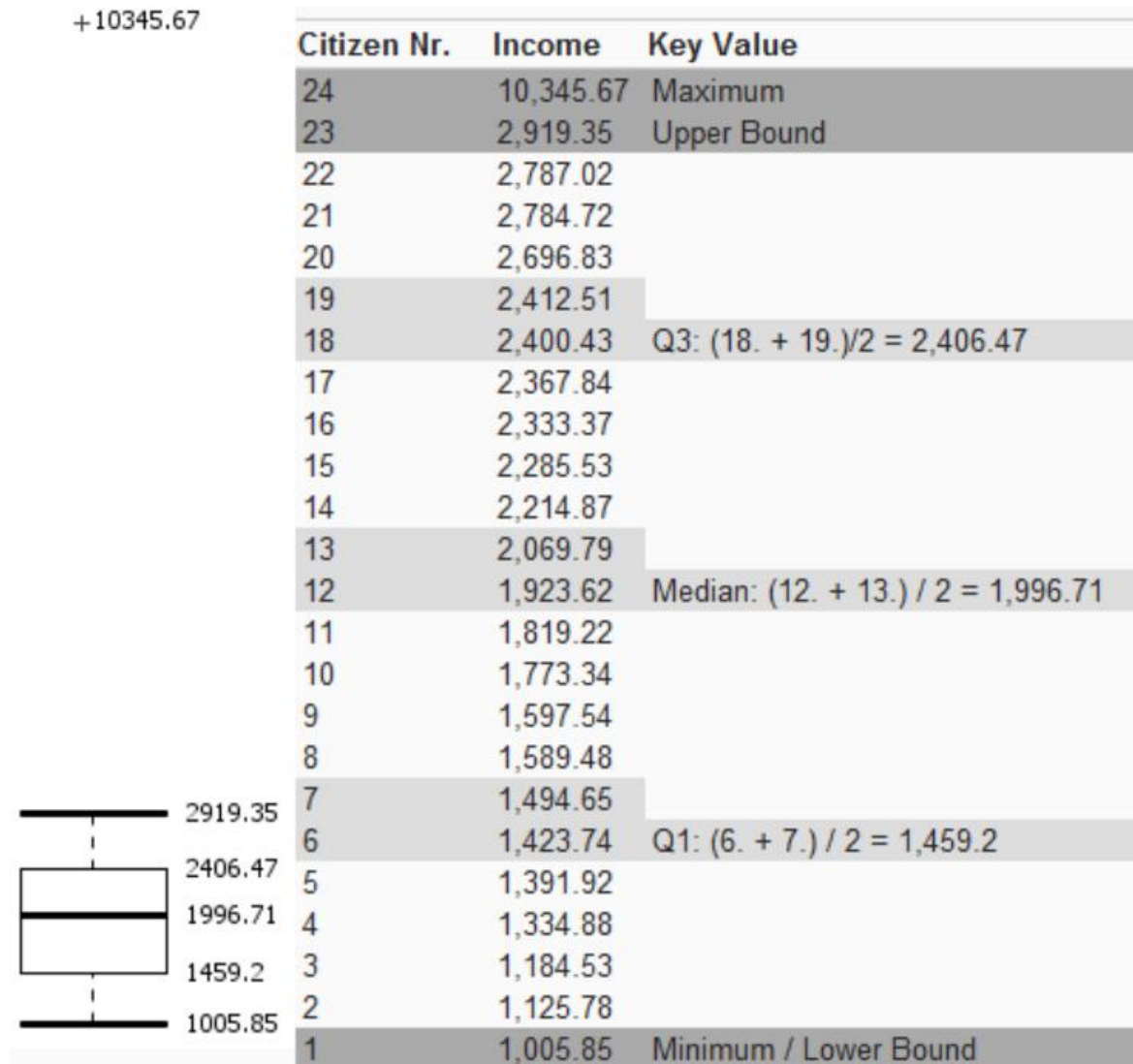
- **Diagrama de Caja y Bigotes**

- Resumen estadístico de la distribución y dispersión de una variable
- El tope de la caja representa el Cuartil 3 (los valores inferiores comprenden el 75% de los datos), y la base el Cuartil 1. El centro es la mediana o Cuartil 2.



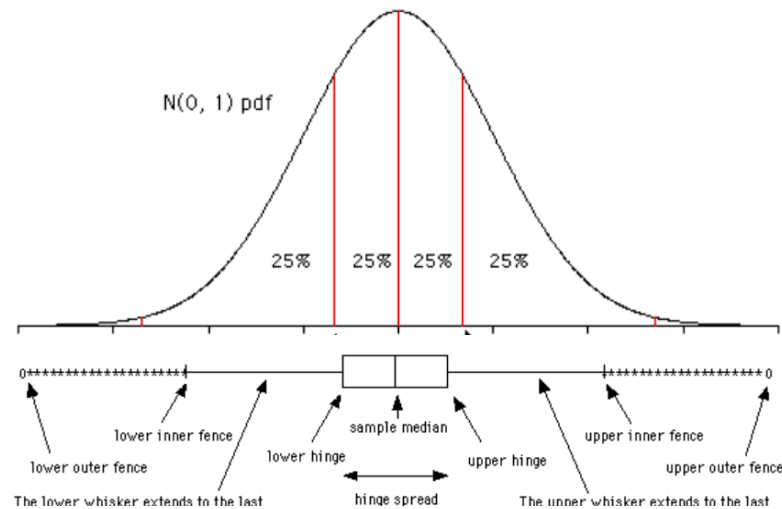
- Los bigotes (barras verticales) representan los límites superior e inferior de los datos, excepto aquellos casos cuyo valor supere:
  - $1.5 \times \text{IQR}$  : atípicos
  - $3 \times \text{IQR}$ : outliers

## 2.2 Datos de 2 y 3 dimensiones



## 2.2 Datos de 2 y 3 dimensiones

- Cuando la mediana está por el centro de la caja y los bigotes tienen una extensión similar, se dice que los datos son más o menos **simétricos (distribución normal)**.



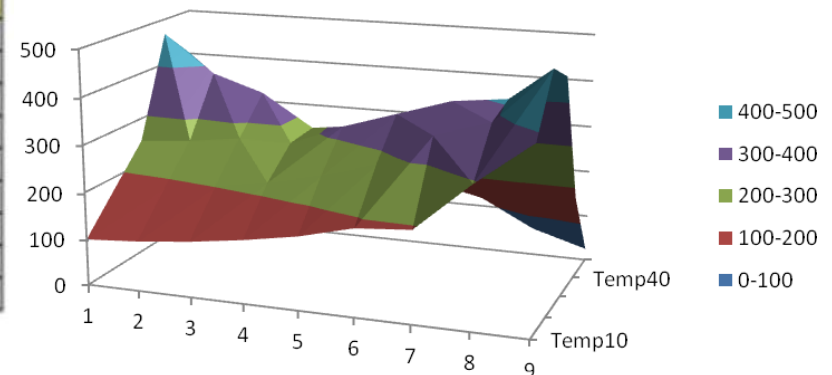
- Entre la mediana y un cuartil, hay el mismo número de casos (25%) que entre la mediana y el otro cuartil; por eso, cuanto más pegada esté la mediana a un cuartil significa que la dispersión de datos en ese rango de valores es menor.

## 2.2 Datos de 2 y 3 dimensiones

- **Diagrama de Superficie**

- Representar 3 variables numéricas
- La variable dependiente se muestra tanto en el eje Z como con colores sobre la superficie.
- Ejemplo excel que, en función de la *temperatura* y el *tiempo*, indica *resistencia* de un cuerpo data un tensión fija.

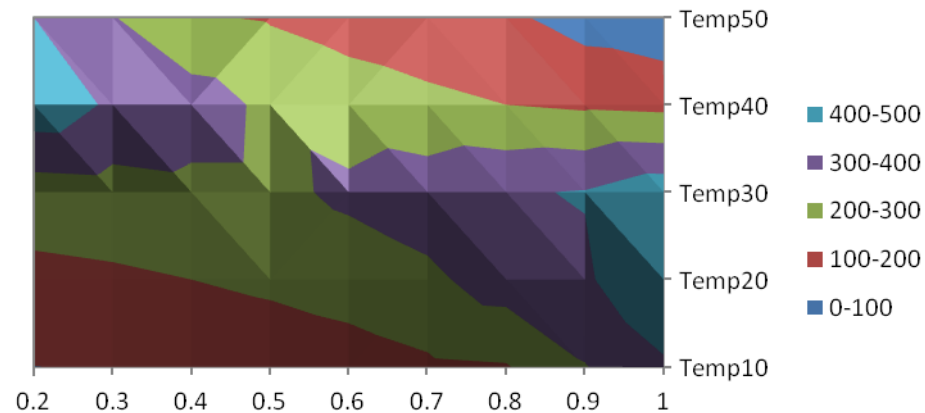
Temperatura Segundos	10	20	30	40	50
0,2	99	175	250	467	400
0,3	107	185	260	385	305
0,4	119	200	275	349	209
0,5	135	220	275	279	192
0,6	155	245	320	245	163
0,7	184	279	356	220	144
0,8	193	349	392	200	118
0,9	295	385	405	185	59
1,0	384	499	459	175	25



## 2.2 Datos de 2 y 3 dimensiones

- **Diagrama de Contorno**

- Se aprovecha que los colores representan valores, y así la representación es en solo 2 dimensiones.
- Es como si el diagrama de superficie se observara desde arriba

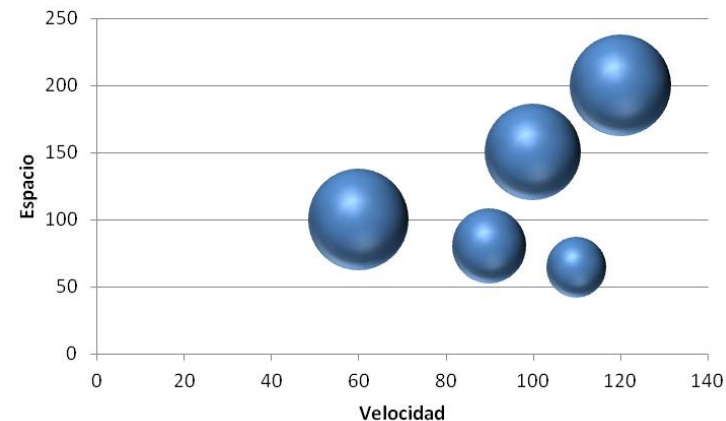


## 2.2 Datos de 2 y 3 dimensiones

- **Diagrama de Burbujas**

- 3 variables numéricas: ejes X-Y, y el tamaño de la burbuja es proporcional a la tercera variable.
- Una burbuja por registro: no sirve para bases de datos con muchos casos
- ó una burbuja por valor de una cuarta variable categórica

Velocidad	Espacio	Tiempo
80	100	1.25
90	80	0.88888889
100	150	1.5
110	64	0.58181818
120	200	1.66666667



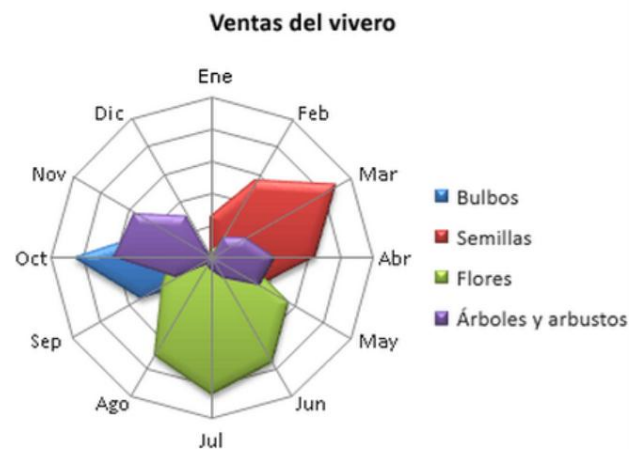


## 2.2 Datos de 2 y 3 dimensiones

- **Diagrama radial o de estrella**

- Círculo dividido en tantos tramos como valores tenga una variable categórica X
- Una estrella por valor de una variable categórica Y
- La estrella tiene una esquina por cada valor de X que aparece en la base de datos simultáneamente a Y
- El alcance de cada esquina, desde el centro, es proporcional a una variable numérica Z
- → 2 variables categóricas y 1 numérica
  - Ejemplo de Office:

- X → mes del año
- Y → tipo de venta
- Z → frecuencia de ventas



## 2.3 Datos multivariados

- Registros representados  $n$  variables, siendo  $n$  cualquier valor mayor que 3. si es 3 ó menos, conviene utilizar las técnicas anteriores ya que son más intuitivas.
- 4 categorías de Técnicas de Visualización para datos multivariados de  $n$  dimensiones y  $m$  registros:
  - **Proyección Geométrica:** proyección de esas  $n$  dimensiones a un espacio bidimensional; u organización de las dimensiones en pares de dos.
  - **Basadas en Píxeles:** habrá  $n$  marcos y  $m$  píxeles en cada marco, donde el color del documento representado por un pixel dependerá del valor de la variable que representa el marco.
  - **Jerárquicas:** el espacio bidimensional donde se realiza la representación se particiona en tantos subespacios como sean necesarios para representar las  $n$  dimensiones.
  - **Basadas en Iconos:** las variables se presentan con pictogramas

## 2.3.1 Datos multivariados – Proyección Geométrica

### PROYECCIÓN GEOMÉTRICA

- Se **transforma el espacio** multivariado a otro de menor dimensiones:
  - Espacio cartesiano
  - Espacio donde las coordenadas se representan de forma no convencional
- Muy **eficiente** cuando se dispone de muchas variables.
- El **orden** de las variables afecta a la representación
- Los registros pueden aparecer **superpuestos** en la representación
- **Técnicas:**
  - Matriz de dispersión
  - Matriz de proyecciones
  - HyperSlice
  - Hyperbox
  - Coordenadas Paralelas y Paralelas Circulares
  - Curva de Andrew
  - Table Lens

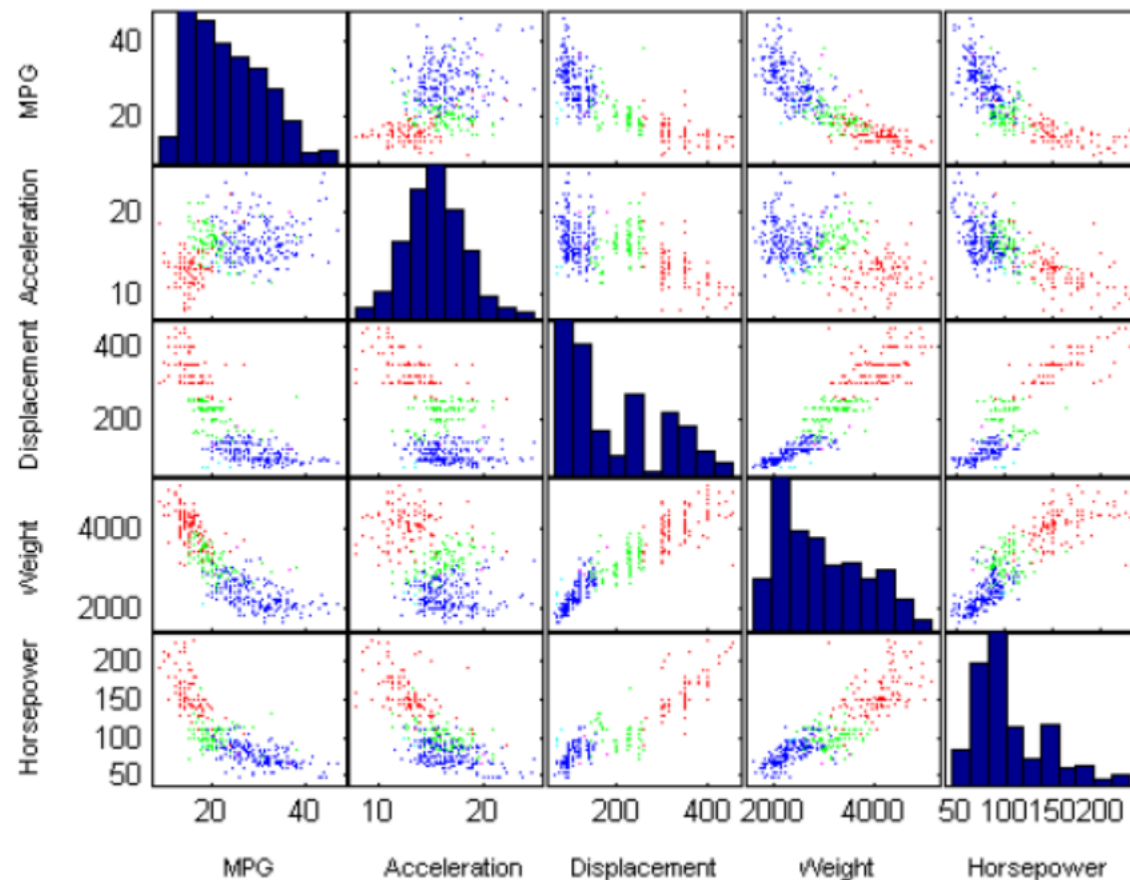
## 2.3.1 Datos multivariados – Proyección Geométrica

### Matriz de dispersión

- Un diagrama de dispersión por cada **par de variables**, formando una matriz
- La **diagonal principal** representa la distribución de cada variable. Si la variable es discreta, entonces es un histograma.
- **Cada punto** dentro de un diagrama representa un registro:
  - Poco eficiente cuando la base de datos es muy grande
- El color de cada punto puede representar el valor de otra variable (categorizada): **brushing**
- Ejemplo:
  - Base de datos con 6 variables:
    - MPG (Miles per Galon)
    - Aceleración
    - Displacement (volumen de aire por ciclo del motor)
    - Peso
    - Horsepower
    - Cilindros

## 2.3.1 Datos multivariados – Proyección Geométrica

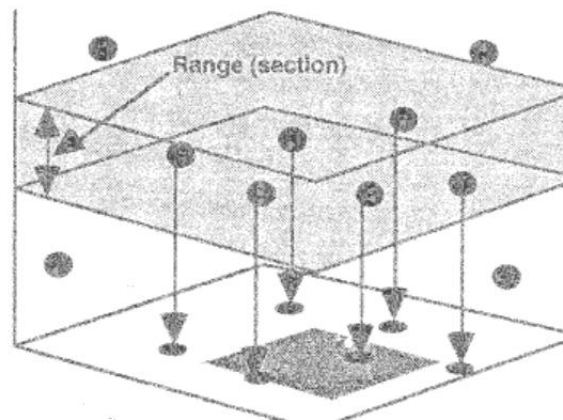
- Hacemos una matriz de dispersión para las primeras 5 variables
- El número de cilindros se discretiza en 3 intervalos de valores, a los que se les



## 2.3.1 Datos multivariados – Proyección Geométrica

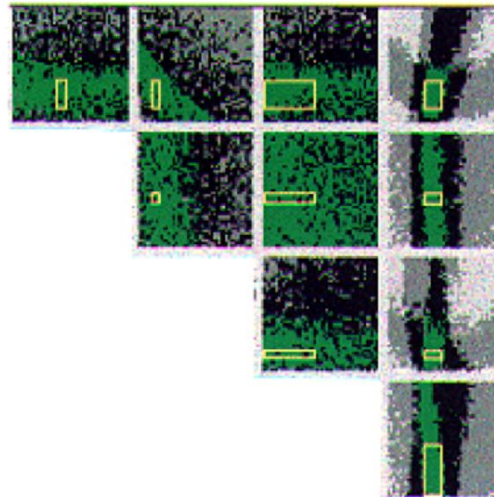
**Matriz de proyecciones**[Furnas and Bujá. Projection Views: Dimensional Inference through Sections and Projections. Journal of Computational and Graphical Statistics. Vol 3, N 4. 1994]

- Matriz de proyecciones por cada par de las  $n$  variables
- Para cada par de variables, una proyección es una **sección+ proyección** de otra variable de interés:
  - Sección: definición de un **rango de valores** en el que queremos que caiga la variable de interés
  - Proyección: proyección a 2D de **únicamente los registros que cumplan** las condiciones de la sección, donde el color del punto depende del grado de satisfacción de las condiciones indicadas para la sección



## 2.3.1 Datos multivariados – Proyección Geométrica

- La diferencia práctica con la matriz de dispersión es que en este caso no se proyectan todos los registros sino solo los que cumplen los criterios de la sección.
- Rectángulo amarillo: zona de resistencia a diferentes valores; es decir, los registros que aún se seguirían proyectando aunque la sección cambiara mucho: **enhancement**



## 2.3.1 Datos multivariados – Proyección Geométrica

**HyperSlice** [J.J. van Wijk and R. van Liere. HyperSlice: Visualization of Scalar Functions of Many Variables. Proceedings of the 4th IEEE Conference on Visualization. 1993]

- Variables numéricas
- Hay que **definir una función escalar**  $f$  tal que

$$f(x_1, x_2, \dots, x_n) \rightarrow \mathbf{R}$$

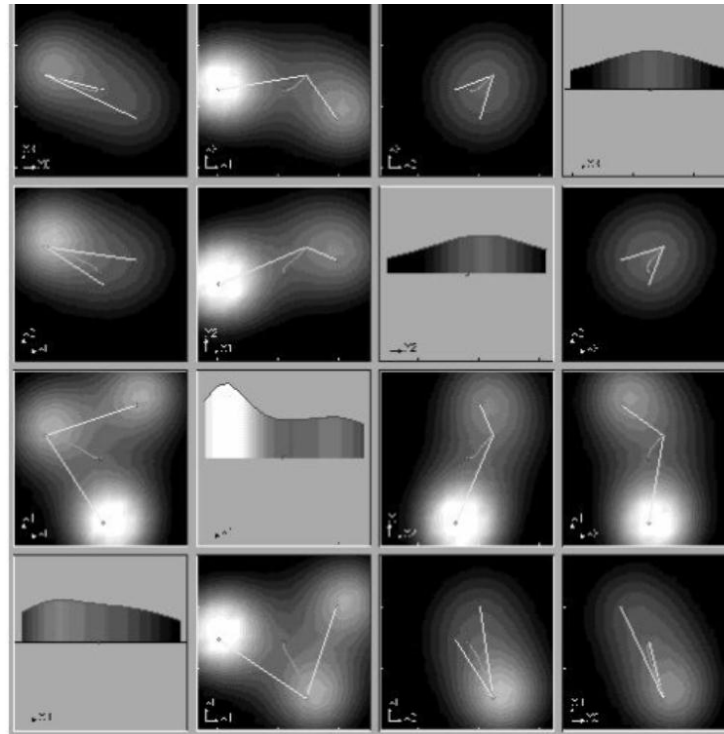
- De nuevo, **matriz** de casillas para cada par de variables
- En cada casilla, la **tonalidad** del color en un punto (i,j) es proporcional al valor de  $f$  para dichos valores, y teniendo un valor fijo del resto de variables.
- La **diagonal principal** representa, para una variable  $x$ , la distribución de  $f$  conforme varía  $x$  y el resto de variables tienen un valor fijo.



## 2.3.1 Datos multivariados – Proyección Geométrica

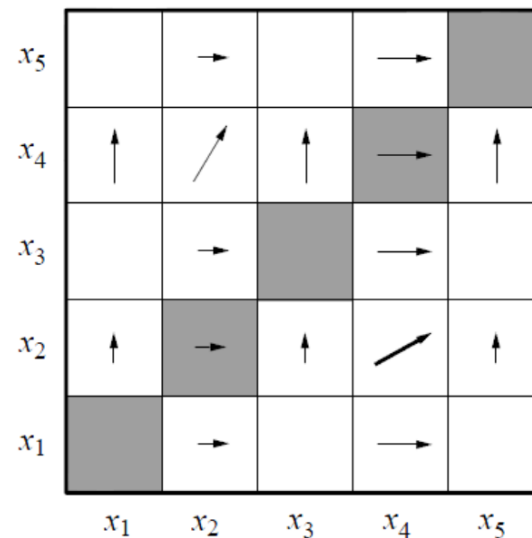
Ejemplo:

- Hay pares de variables (1,3) cuyos rangos de valores no parecen cambiar el valor de la función



## 2.3.1 Datos multivariados – Proyección Geométrica

- El valor fijo de las variables es el valor que tengan en un registro seleccionado.
- Interactuando con el ratón encima de cada casilla, se puede cambiar el registro y ver cómo varían los tonos (valores de la función).

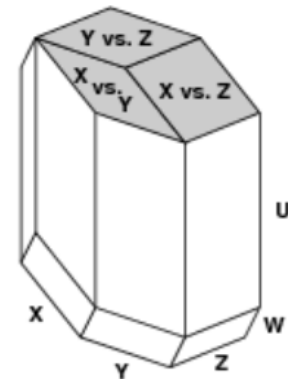


- El camino recorrido por el hyperespacio queda marcado con unas líneas en cada casilla (ver imagen del ejemplo)

## 2.3.1 Datos multivariados – Proyección Geométrica

**HyperBox.** [Alpern and Carter. Hyperbox. Proceedings of the 2nd IEEE Conference on Visualization. 1991]

- Una caja *n-dimensional*.
- Cada cara, como en las matrices, representa un par de variables.
- La caja tendrá  **$n^2$  segmentos y  $n(n-1)/2$  caras.**
- Al dibujar sobre una superficie de 2D, **no todas las variables se tratan igual** (el orden aquí es aún más importante que en las técnicas anteriores) ya que no todas las caras son iguales:
  - Se pueden enfatizar variables
- Un **conjunto direccional** es un conjunto de segmentos con la misma dirección
- Se asigna un conjunto direccional a cada variable



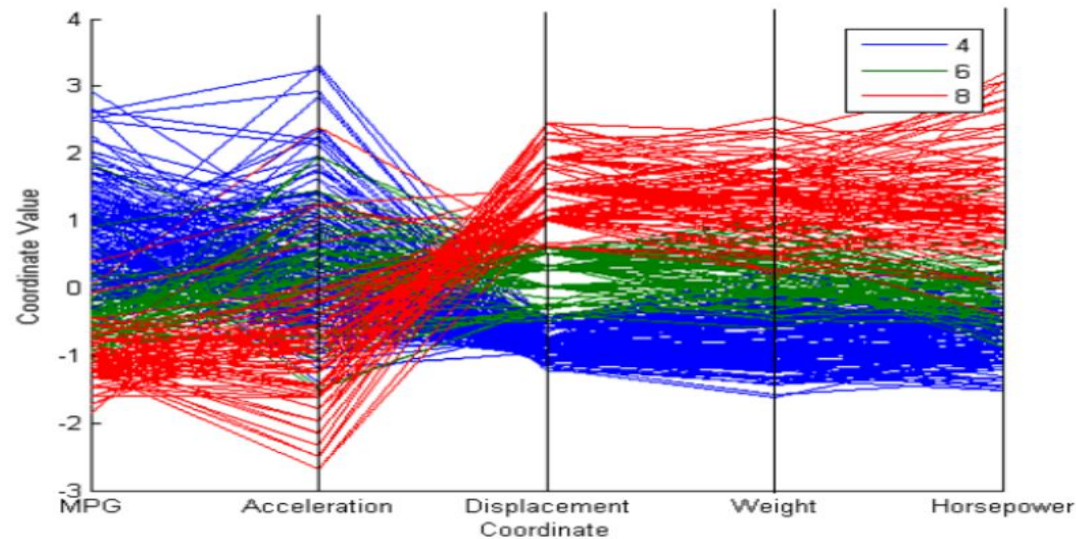
## 2.3.1 Datos multivariados – Proyección Geométrica

- **Coordenadas Paralelas** [A. Inselberg. “Multidimensional Detective”. Proceedings of the IEEE Symposium on Information Visualization. 1997]
- $n$  ejes verticales equidistantes, uno para cada variable y dividido en los posibles valores que la variable puede tomar.
- Hay que escalar los posibles valores en cada eje para que todos los ejes tengan la **misma altura** (escala).
- Cada **registro** se representa por una **línea poligonal** que corta cada eje vertical por el valor que toma la variable correspondiente en dicho registro.

Volvemos al ejemplo de los cilindros:

- MPG (Miles per Galon)
- Aceleración
- Displacement (volumen de aire por ciclo del motor)
- Peso
- Horsepower
- Cilindros (usando brushing)

## 2.3.1 Datos multivariados – Proyección Geométrica



- Muy efectivo para mostrar relaciones entre atributos
- Problemas:
  - El orden de las variables es extremadamente importante para encontrar patrones
  - Al tener muchas variables, se necesita mucho espacio horizontal

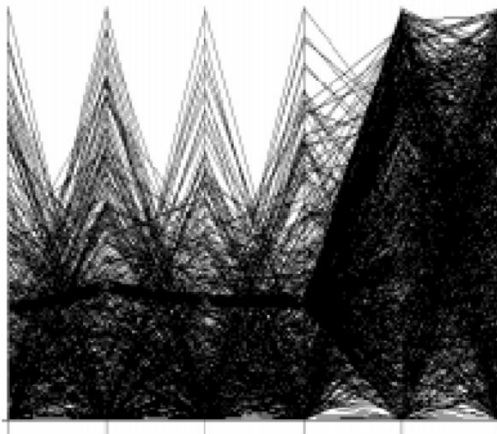
## 2.3.1 Datos multivariados – Proyección Geométrica

- Representación con 15000 registros

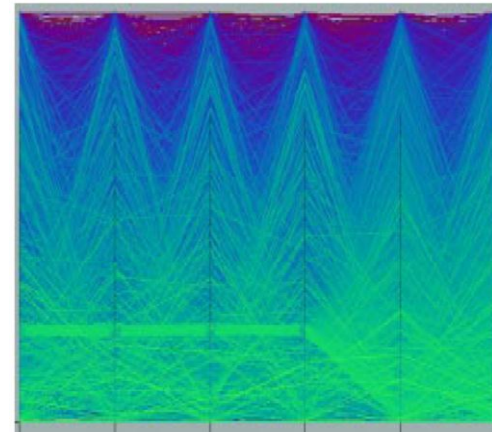


- Soluciones:

- (1) Selección de registros

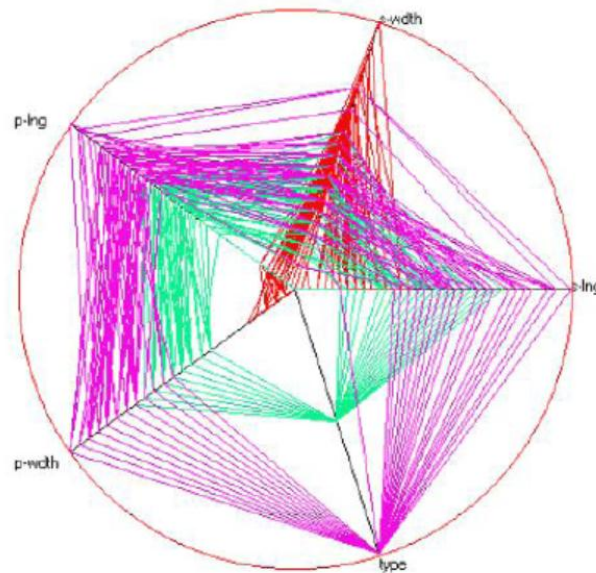


- (2) Brushing



## 2.3.1 Datos multivariados – Proyección Geométrica

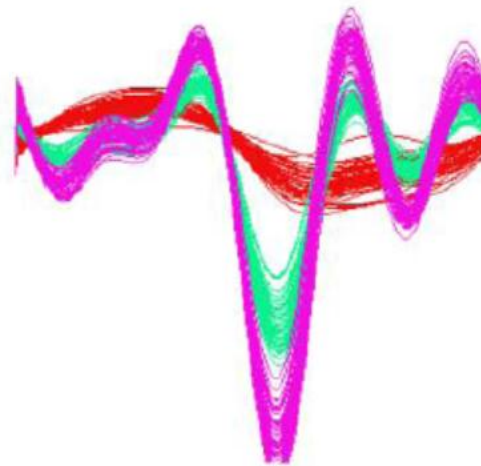
- **Coordenadas Paralelas Circulares** [P.E.Hoffman. Table Visualizations: A Formal Model and Its Applications. Doctoral Dissertation, University of Massachusetts at Lowell. 1999]
- Igual que las anteriores pero los ejes parten del centro de un círculo
- Ejemplo (usando brushing):





## 2.3.1 Datos multivariados – Proyección Geométrica

- **Curva de Andrew** [D.F. Andrews. Plots of High-Dimensional Data". Biometrics. 1972]
- Cada registro se transforma en una línea curva, mediante una transformación matemática (ej, transformada de Fourier).
- Registros con valores similares en sus variables forman curvas similares:
  - detección de clusters
  - detección de outliers
- Ejemplo (usando brushing):





## 2.3.1 Datos multivariados – Proyección Geométrica

- **Table Lens** [R. Fao and S. K. Card, “The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information”, *Proceedings of the SIGCHI Conference on Human Factors in Computer Systems: Celebrating Interdependence. 1994*]
- Utiliza una distribución de los datos que nos resulta muy familiar: **tabla con los registros en filas y las variables en columnas**
- Pero cada posición (i,j) no es un valor numérico sino una representación **gráfica** de dicho valor, en forma de **columna horizontal** (si es valor real) **o histograma** (si es categórico).
- Se puede **interaccionar**, pero esto lo veremos en los temas de Interacción y Distorsión.

