Análisis Inteligente de Datos

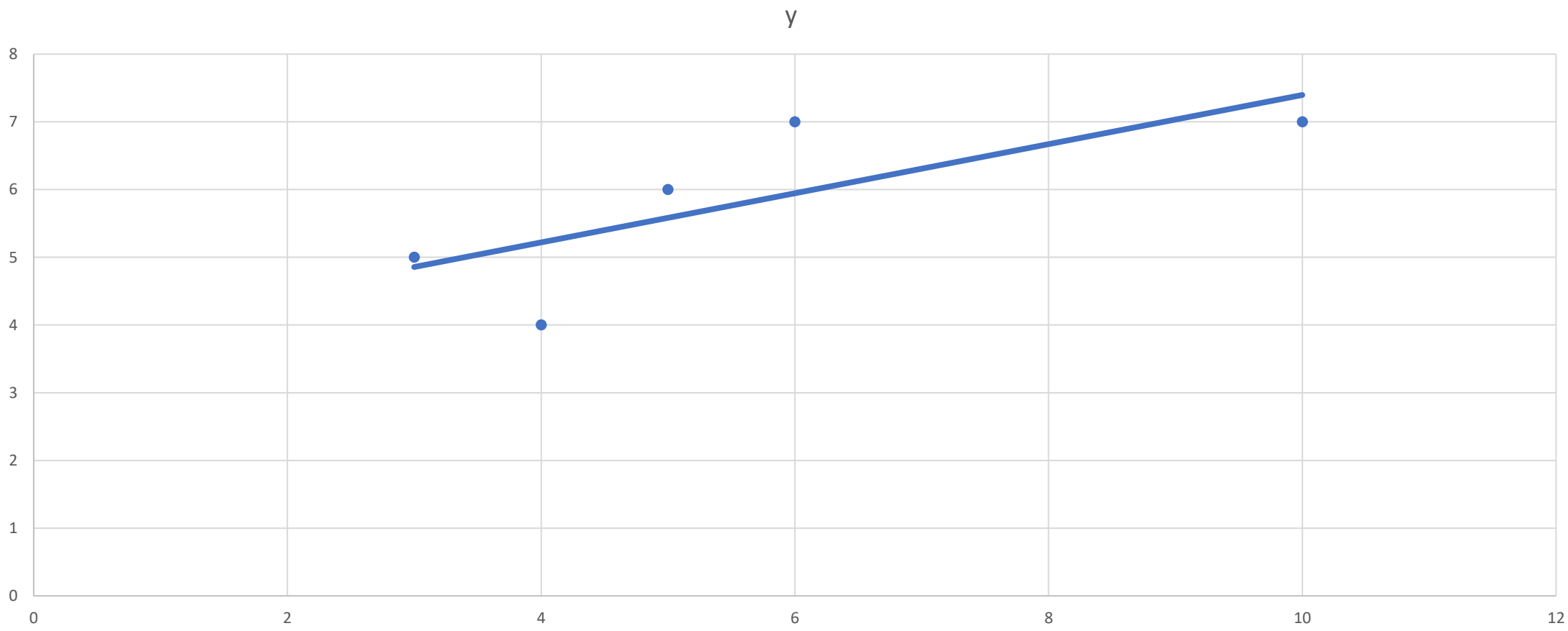
Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Profesora: Mónica Cantoni

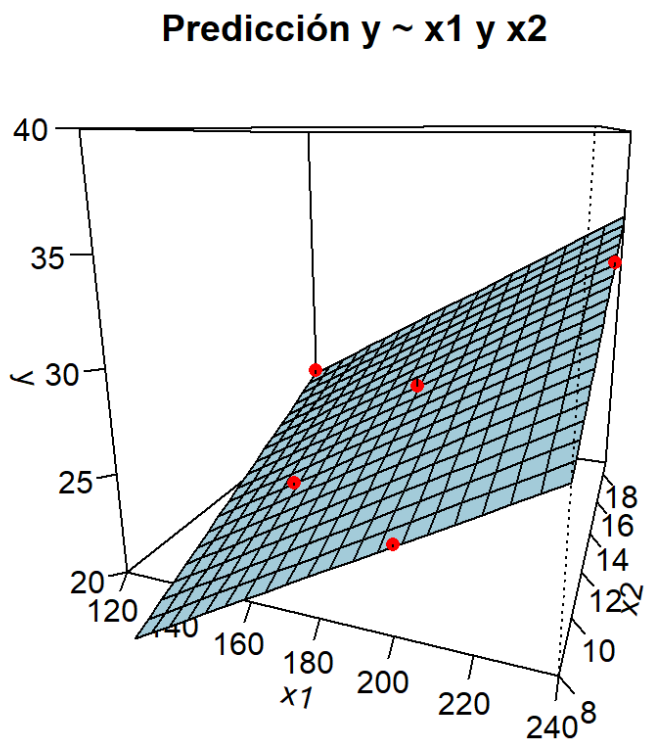
Ayudantes: Cecilia Oliva - Fabiana Rossi - Pamela Pairo

Clase 05 – Regresión y Análisis Multivariado

Regresión y Correlación Lineal Simple



Regresión Múltiple



Información multivariada

	Variable 1	Variable j		Variable p
Individuo 1	$X_{1,1}$	$X_{1,j}$	…	$X_{1,p}$
Individuo i	$X_{i,1}$	$X_{i,j}$	…	$X_{i,p}$
⋮	⋮	⋮		⋮
Individuo n	$X_{n,1}$	$X_{n,j}$	…	$X_{n,p}$

$$X = \begin{pmatrix} X_{1,1} & X_{1,j} & X_{1,p} \\ X_{i,1} & X_{i,j} & X_{i,p} \\ X_{n,1} & X_{n,j} & X_{n,p} \end{pmatrix}$$

Objetivos del análisis exploratorio

Algunos de los objetivos que se fijan en el análisis exploratorio son los siguientes:

Conocer los datos.

Descubrir regularidades.

Verificar la existencia de estructuras ocultas.

Entender los patrones descubiertos.

Resumir información.

Hallar asociaciones de variables.

Detectar anomalías

Tabla de clasificación cruzada

Se han tabulado las consideraciones respecto del ambiente laboral y el salario, que tienen 1441 empleados de diferentes empresas que participaron de un estudio. Se presenta la distribución conjunta de estas dos variables.

	Ambiente laboral bueno	Ambiente laboral malo	Total
Salario bueno	258	280	538
Salario malo	184	719	903
Total	442	999	1441

Gráfico de mosaicos

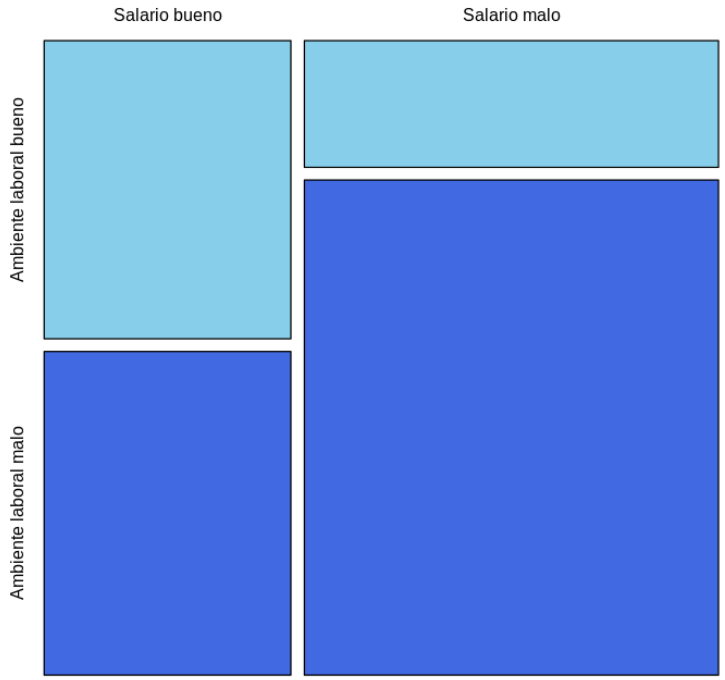


Diagrama de dispersión

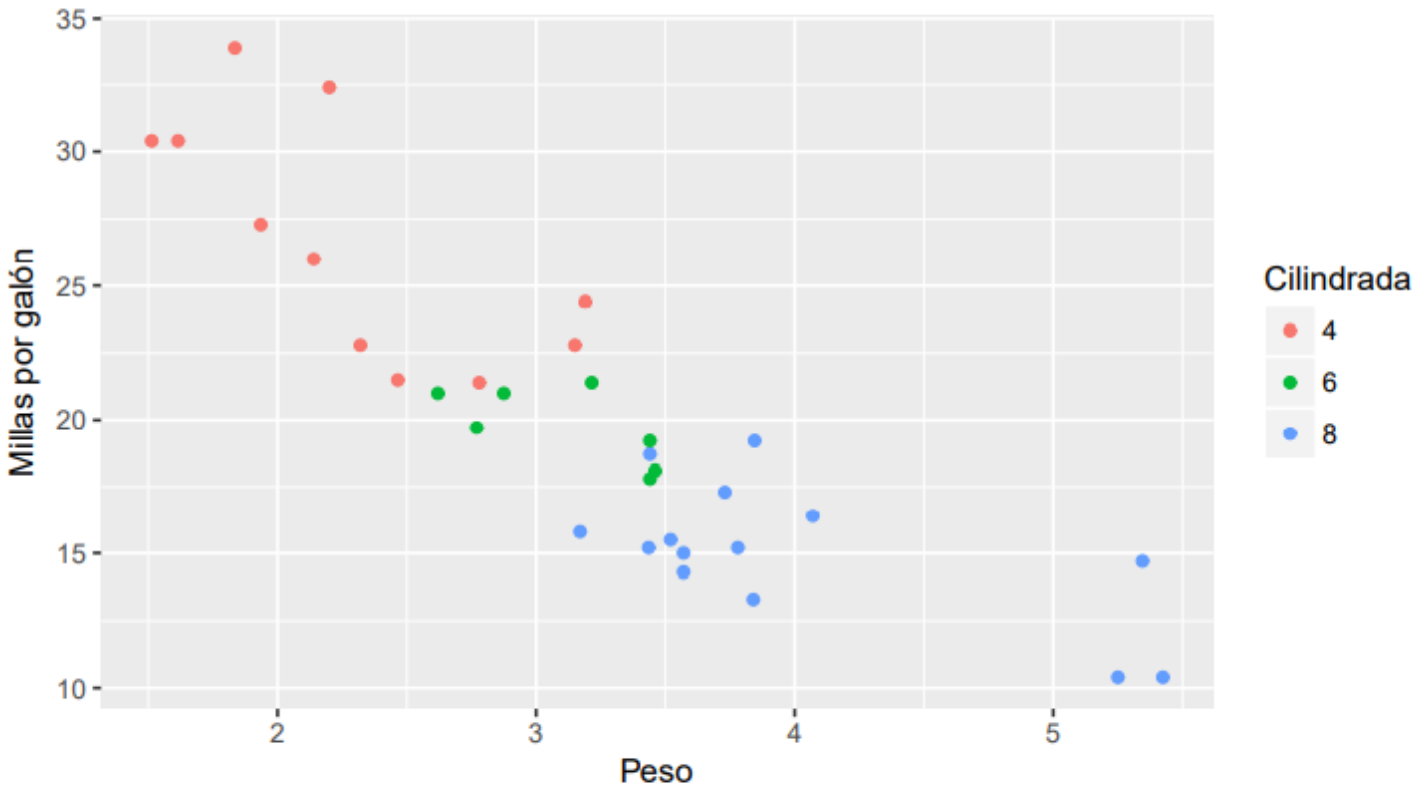


Figura 2.25: Diagrama de dispersión para tres poblaciones

Dispersograma

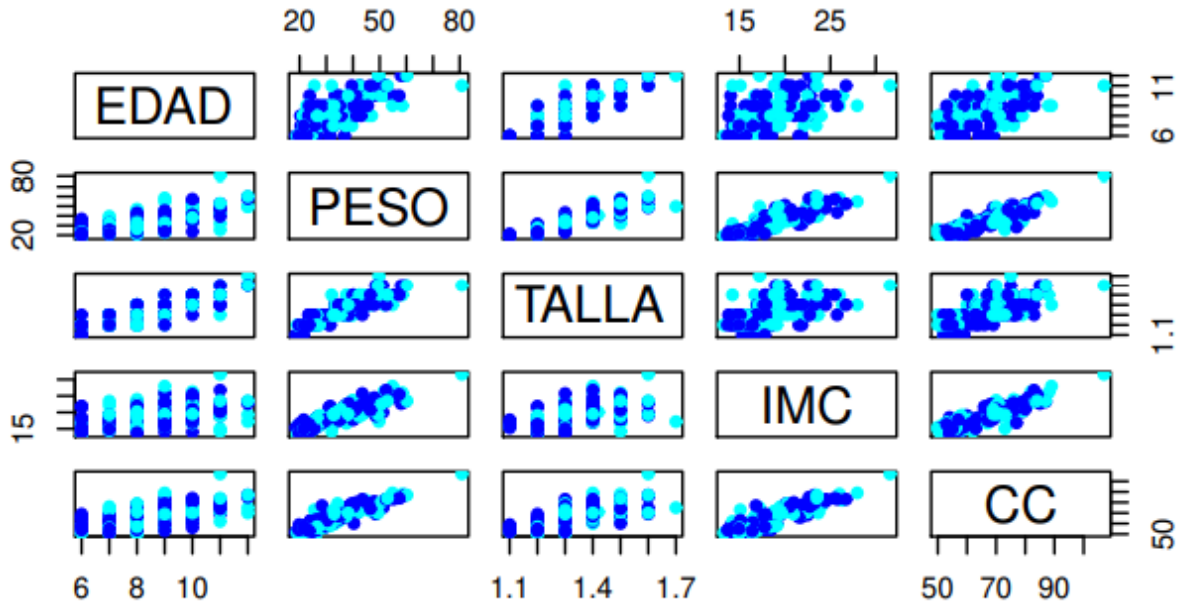


Figura 2.26: Dispersograma

Gráfico de coordenadas paralelas

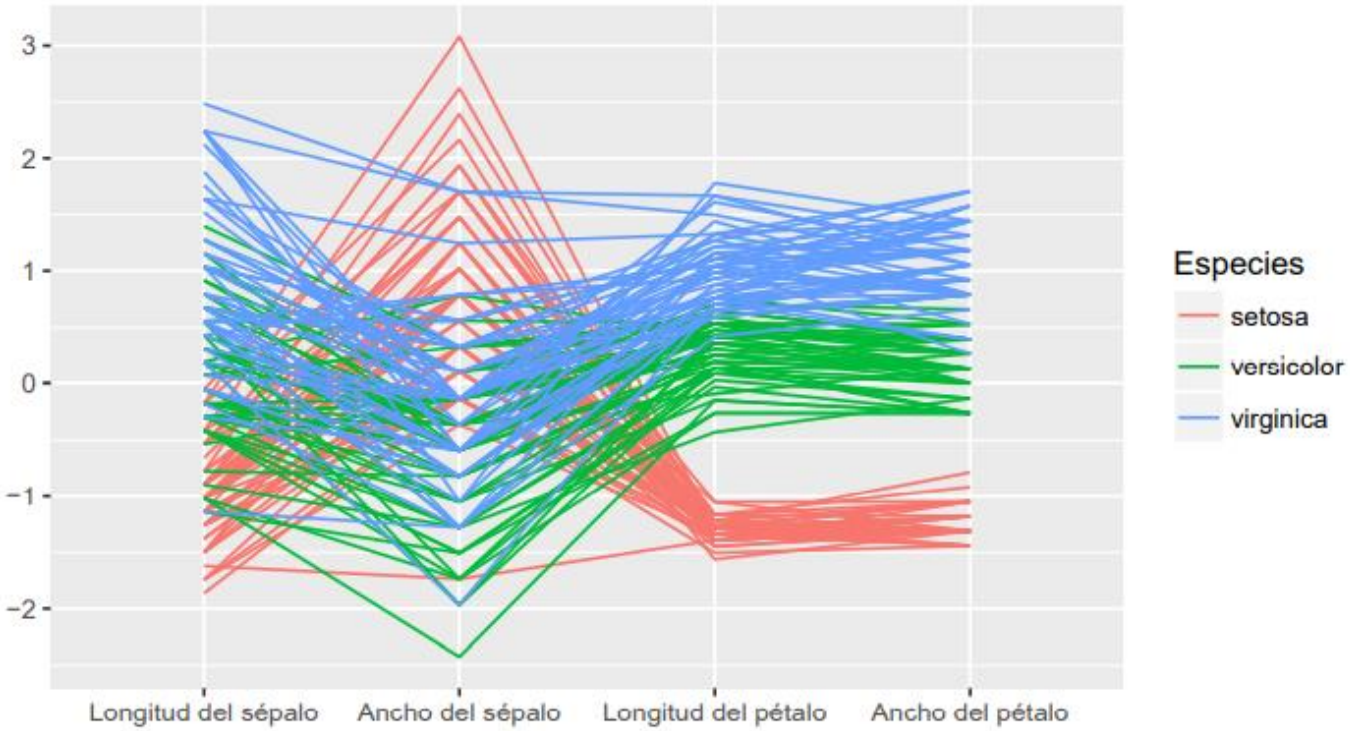


Figura 2.27: Gráfico de coordenadas paralelas

Gráfico de perfiles multivariados

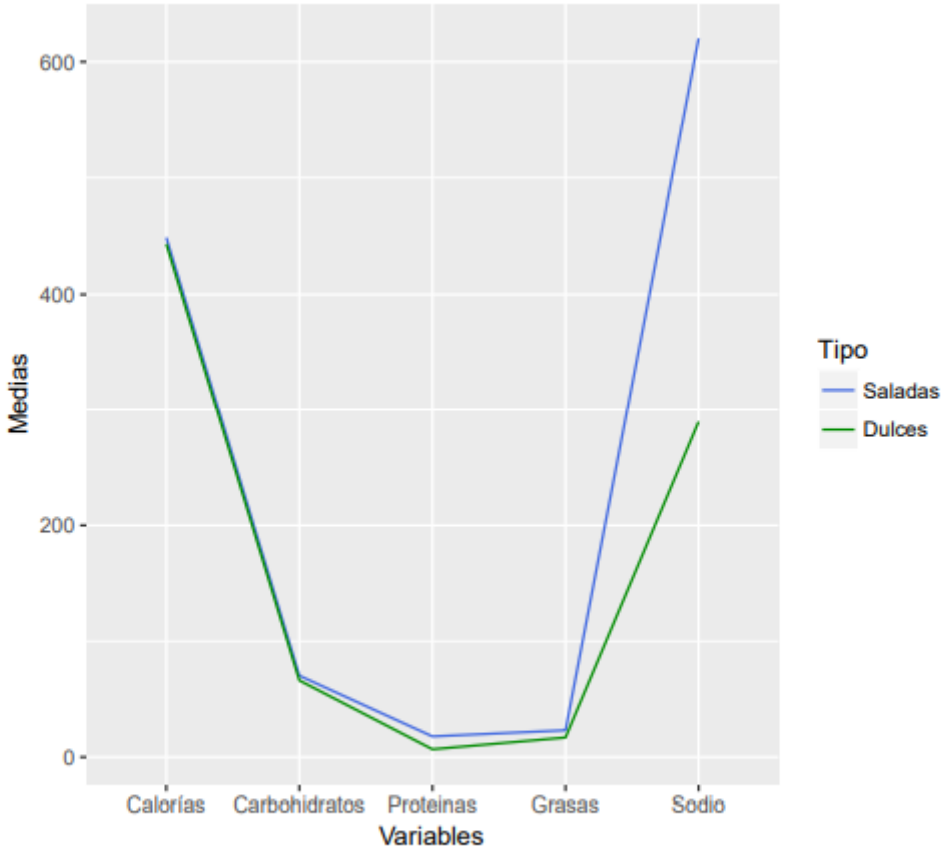
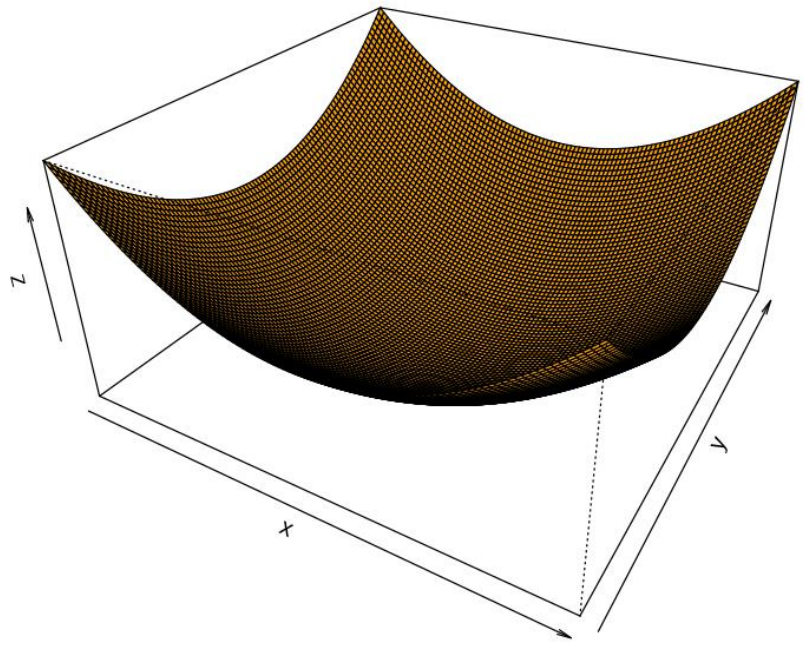


Figura 2.28: Gráfico de perfiles

Gráfico en 3D y Curvas de nivel

Paraboloide



$z = x^2 + y^2$

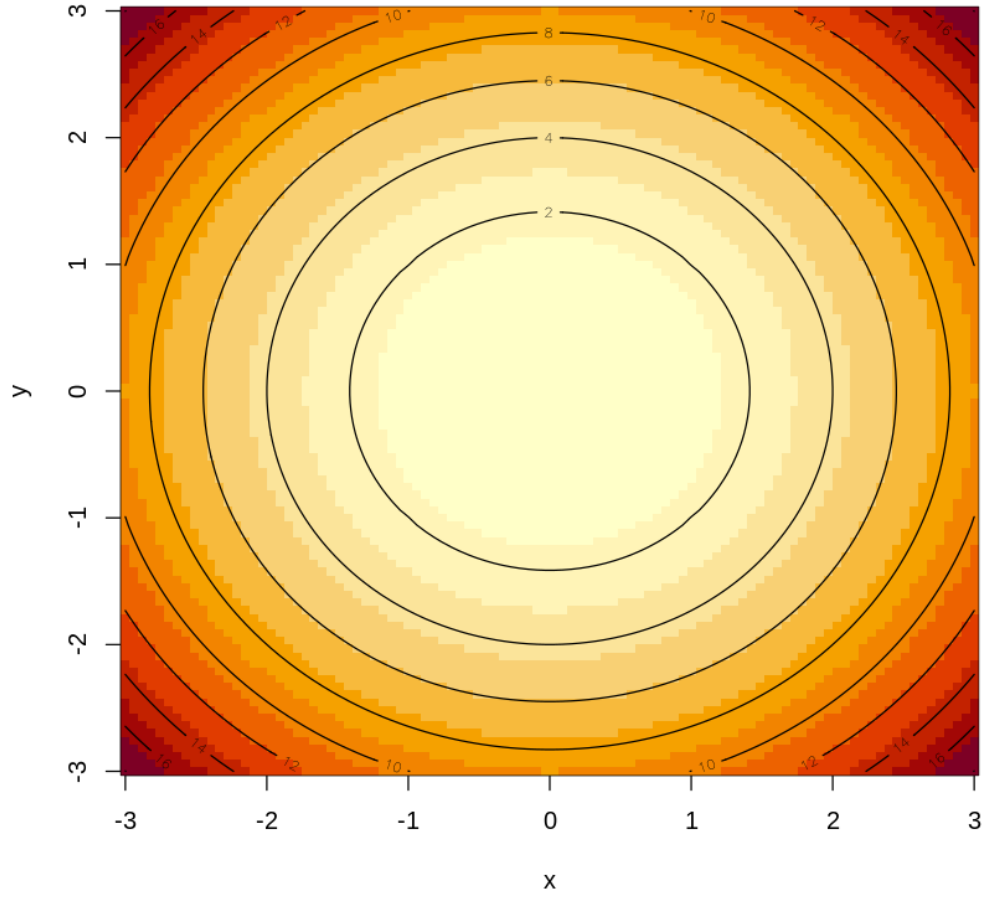
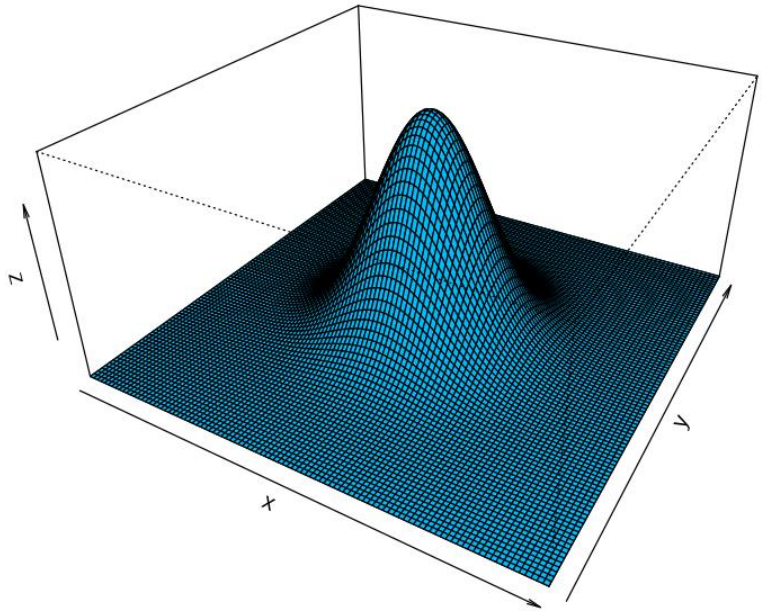
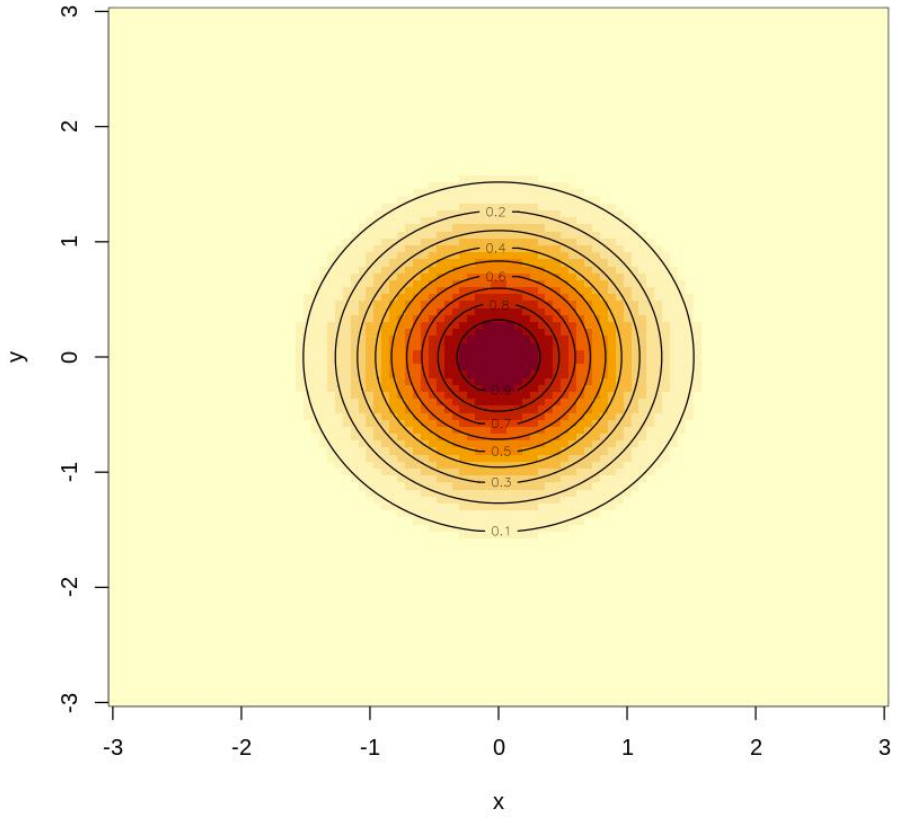
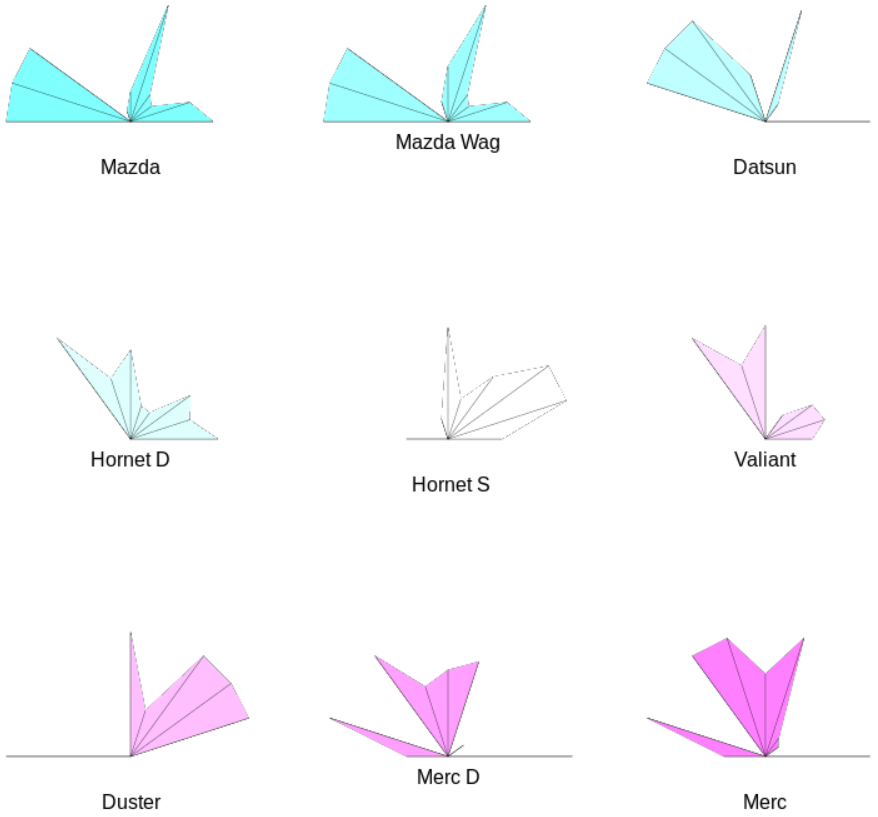


Gráfico en 3D y Curvas de nivel

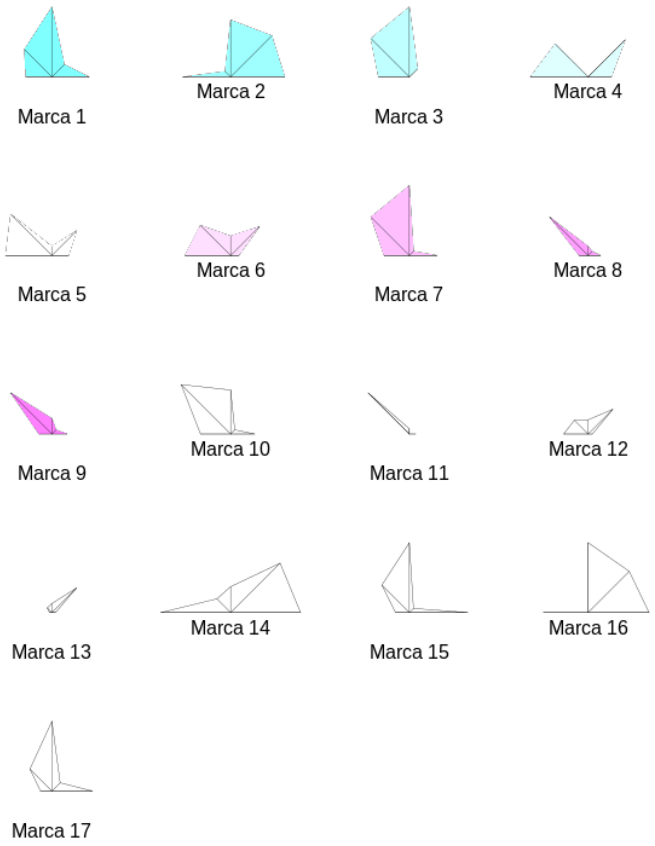




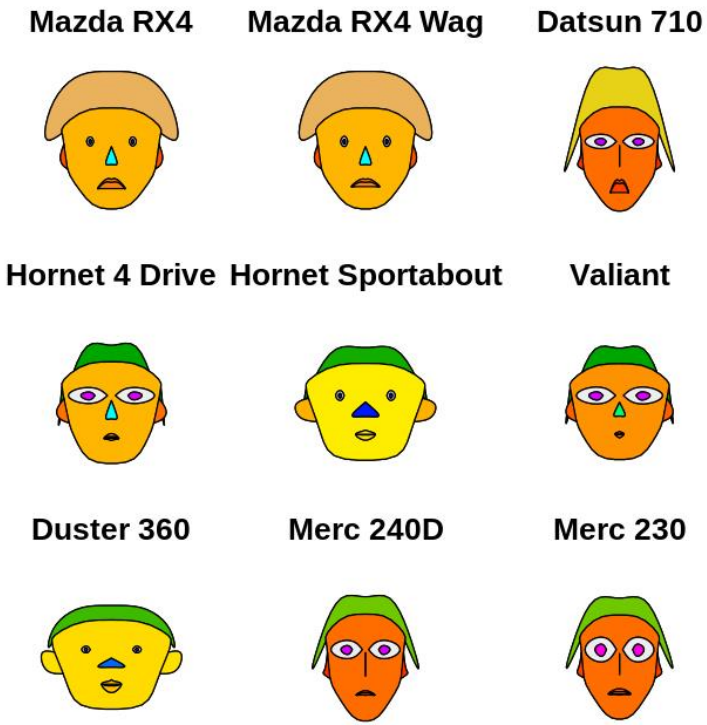
Gráficos de estrellas

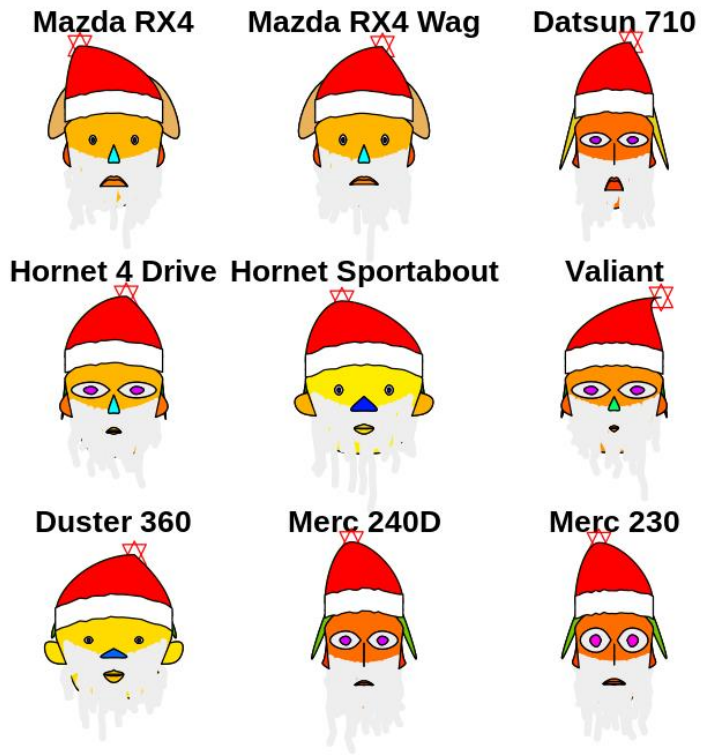


Gráficos de estrellas



Gráficos de caras de Chernoff





Transformación del conjunto de datos

- En algunas ocasiones, para optimizar el análisis de la información disponible, es conveniente realizar transformaciones a los datos. Las transformaciones pueden ser por filas o por columnas, o sea por individuos o por variables, dependiendo de los objetivos de las mismas. Los objetivos más usuales de estas transformaciones son:
- hacer comparables las magnitudes,
- modificar la escala de medición,
- satisfacer alguna propiedad estadística.

Transformaciones por variables

Variables aleatorias estandarizas

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Transformaciones por individuos

Variables aleatorias estandarizas

$$T(x) = \frac{x - \bar{x}}{x_{máx} - \bar{x}} \text{ si x} > \bar{x}$$

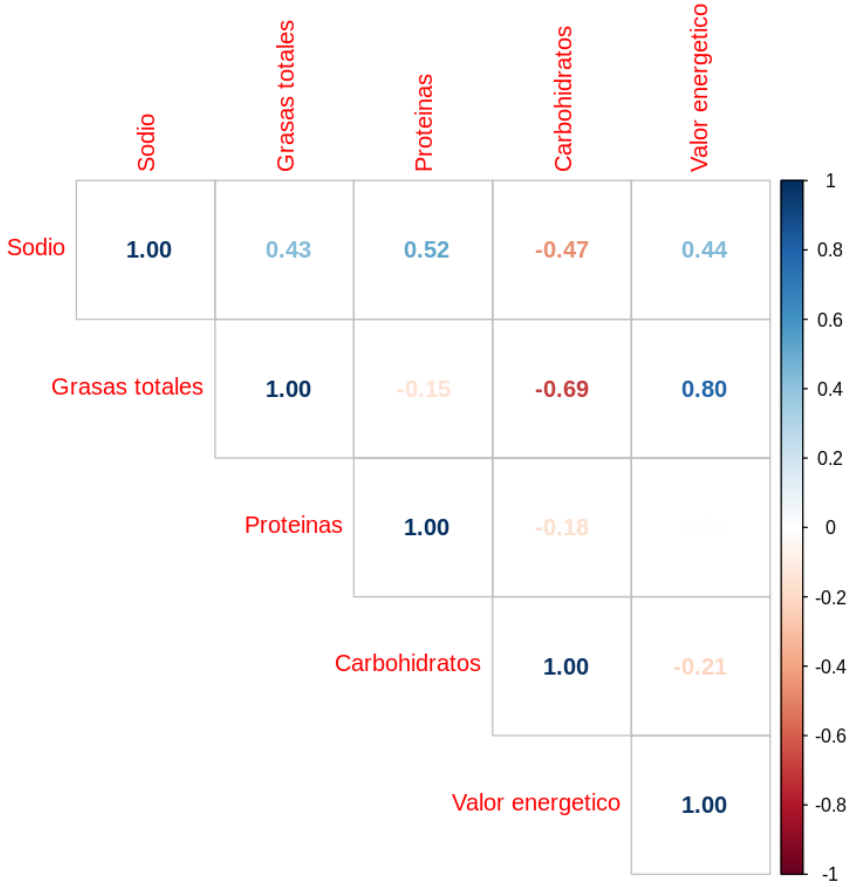
$$T(x) = \frac{x - \bar{x}}{\bar{x} - x_{min}} \text{ si x} < \bar{x}$$

Las puntuaciones superiores a la media resultan positivas y las que resulten inferiores a la media resultarán negativas

Covarianza y Correlación

- Covarianza es una medida de asociación lineal entre dos variables.
- Correlación es una medida de asociación lineal definida como la covarianza de los datos estandarizados.
- La correlación mide el grado de asociación lineal entre dos variables.
- Toma valores entre -1 y 1.

Correlograma



Alternativas robustas para posición y escala

Las estadísticas robustas proponen métodos similares a los de la estadística clásica, pero que no se vean afectados por la presencia de observaciones atípicas (outliers en inglés) u otras desviaciones de los supuestos de un modelo.

Por lo general las observaciones atípicas en bases grandes de datos no pueden ser eficientemente detectadas analizando por separado cada variable. La detección resulta más eficiente estudiando el conjunto general de todas las variables.

Los outliers, en casos multivariados, pueden provocar dos tipos de efectos:

El efecto de enmascaramiento se produce cuando un grupo de outliers esconden a otro/s. Es decir, los outliers enmascarados se harán visibles cuando se elimine/n el o los outliers que los esconden.

El efecto de inundación ocurre cuando una observación sólo es outlier en presencia de otra/s observación/es. Si se quitara/n la/s última/s, la primera dejaría de ser outlier.

Distancia de Mahalanobis

Este concepto fue introducido por Mahalanobis y se diferencia de la distancia euclideana pues considera la correlación entre las variables.

Esta distancia es muy usada en Estadística Multivariada.

Sean X e Y dos variables aleatorias pensadas como vectores columna y con la misma distribución de probabilidad.

Si $\sum$ es la matriz de covarianzas, se define la distancia de Mahalanobis como:

$$d_m(X,Y) = \sqrt{(X-Y)^t \sum^{-1} (X-Y)}$$

Vector de medianas

Sathishkumar y Thangavel proponen sustituir el vector de medias por un vector de medianas y calcular la matriz de covarianza para el conjunto de las k observaciones con menor distancia de Mahalanobis al vector de medianas.

Realizar una estimación robusta de la matriz de covarianzas puede entenderse como estimar la covarianza de una buena parte de los datos.

MVE (Minimum Volume Ellipsoid) (Elipsoide de volumem mínimo)

Es un estimador que minimiza el volumen de la matriz de covarianza asociada a la submuestra, se basa en el elipsoide de menor volumen que cubre h de las n observaciones. Se trata de un equivariante afín, un estimador robusto de alto desglose de la ubicación de múltiples variables y de dispersión.

El MVE puede calcularse mediante un algoritmo de remuestreo. Su sesgo de baja hace que sea útil para la detección de outliers en los datos multivariados, mediante el uso de las distancias robustas basadas en MVE.

MCD (Minimum Covariance Determinant) (Determinante de mínima covarianza)

El estimador determinante de covarianza mínima (MCD) es un estimador multivariable robusto, de ubicación y dispersión.

Puede calcularse de manera eficiente con el algoritmo de FAST-MCD de Rousseeuw y Van Driessen.

Dado que la estimación de la matriz de covarianza es la piedra angular de métodos estadísticos multivariados, el MCD se ha utilizado también para desarrollar técnicas multivariantes robustas y eficientes computacionalmente.

Transformaciones de Box Cox

- Si las muestras son suficientemente grandes, es posible aplicar la distribución Normal, basándonos en el Teorema del Límite Central que tiene un nivel aproximado o asintótico. Este es el caso más usual para *data mining* donde se dispone de mucha información y, en general, la información no satisface el supuesto de normalidad.
- Si se desea aplicar una prueba basada en el supuesto de normalidad y los datos disponibles no la satisfacen, una alternativa viable es aplicar transformaciones de Box & Cox.

Transformaciones de Box Cox

- La transformación de Box-Cox es una transformación potencial que corrige la asimetría de una variable, varianzas diferentes o la no linealidad entre variables. En consecuencia, resulta muy útil para transformar una variable y obtener una nueva que siga una distribución normal.

Familia Box Cox

Las transformaciones de Box-Cox vienen dadas para diferentes valores de λ por la siguiente expresión:

$$\begin{cases} \frac{x^{\lambda}-1}{\lambda} \; si \; \lambda \neq 0 \\ \log x \; \; si \; \lambda = 0 \end{cases}$$

siendo y la variable a ser transformada y
λ el parámetro de transformación.
Sin embargo, las **transformaciones más habituales** se describen en la siguiente tabla:

λ	Transformación
-2	$1/x^2$
-1	$1/x$
-,5	$1/\sqrt{x}$
0	$log\ x$
0,5	$\sqrt{x}$
1	x
2	x^2

Transformaciones de Box Cox

- Si el parámetro de la transformación estimado es **cercano a los valores de la tabla anterior, en la práctica es recomendable utilizar el valor de la tabla** en lugar del exacto, ya que será más fácil de interpretar.

¿Preguntas?