Universidad Ricardo Palma

RECTORADO

PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

Formamos seres humanos para una cultura de paz

II PROGRAMA DE ESPECIALIZACIÓN EN "INTRODUCCIÓN AL DATA SCIENCE"

MÓDULO R - RSTUDIO

SESIÓN 3

Visualización de Gráficos Especiales

Expositor: José Cárdenas Garro

josecardenasgarro@gmail.com

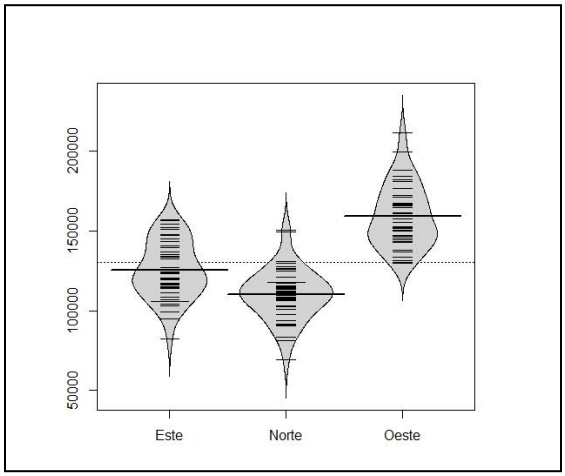
A nuestro recordado Maestro

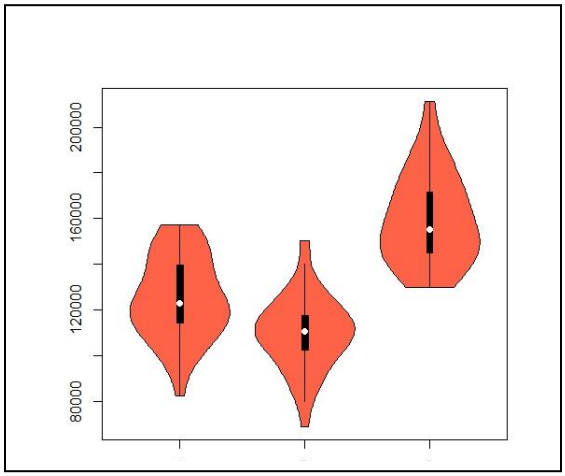
Dr. Erwin Kraenau Espinal, Presidente de la Comisión de Creación de la Maestría en Ciencia de los Datos

PROGRAMA DE ESPECIALIZACIÓN EN “INTRODUCCIÓN AL DATA SCIENCE”

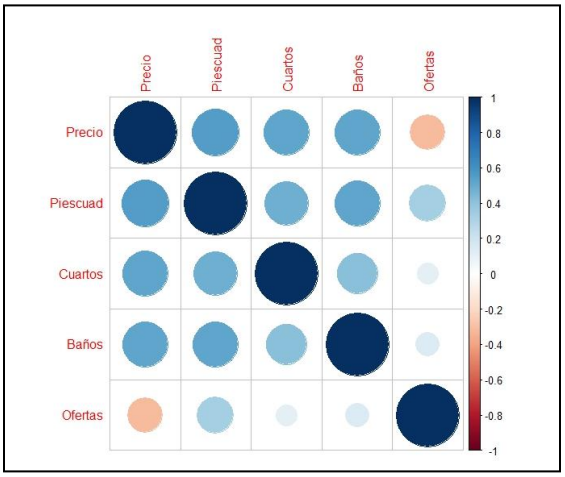
GRÁFICOS ESPECIALES

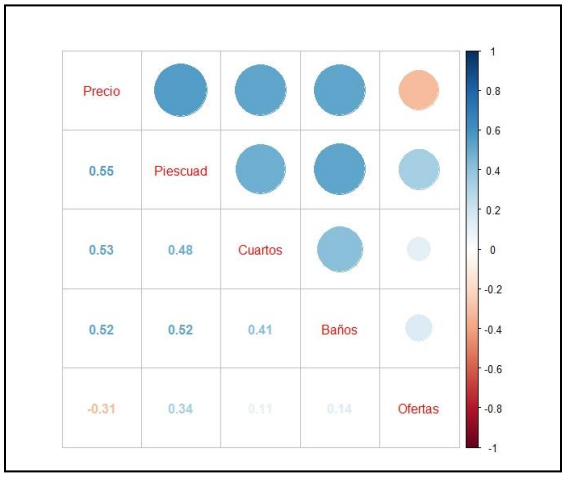
El **beanplot** es una combinación de un gráfico de densidad (doble) con las marcas de todos los datos. Dichas marcas cambian de color si se salen del interior de la doble densidad y se alargan si coinciden algunos datos con el mismo valor.
Para poder comparar los grupos, se señalan las medias de cada grupo y la media general.
Por otra parte, si en la población general hay un factor con dos niveles, se puede considerar un beanplot asimétrico con dos densidades distintas en función del factor.





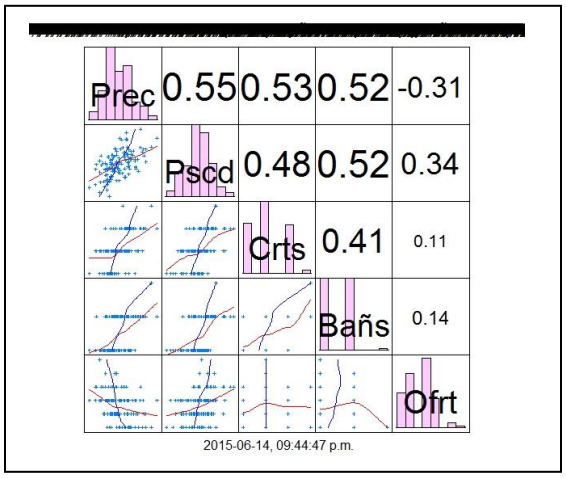
El paquete **corrplot** es un dispositivo gráfico de una matriz de correlación e intervalos confidenciales.

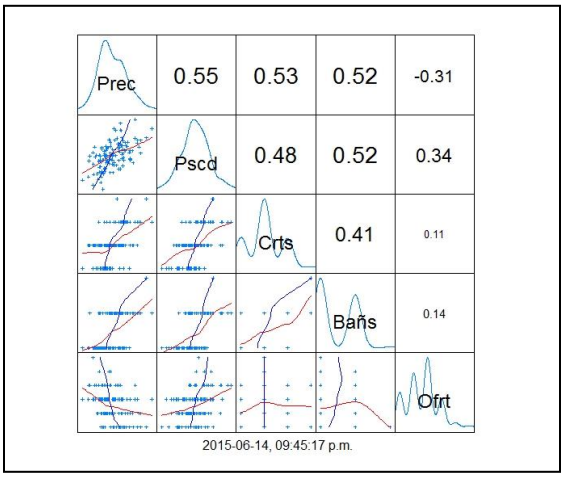




Autor: Dr. Erwin Kraenau - Lic. José Cárdenas

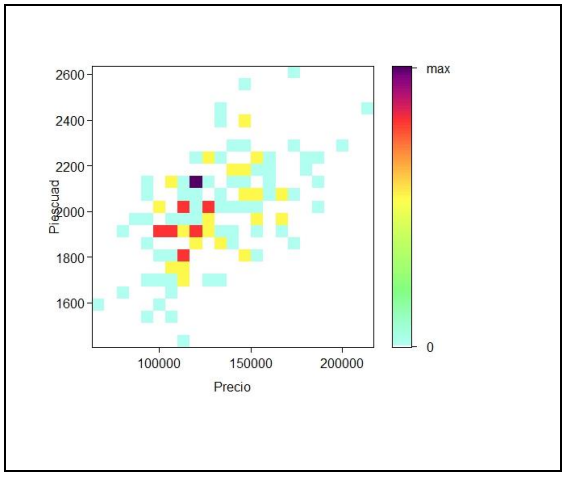
La sentencia **SplomT** crea una matriz de diagramas de dispersión con **a)** covarianzas (con tamaño de la escritura proporcional al tamaño) en el triángulo superior, **b)** histogramas (con suavización) y los nombres de las variables en la diagonal, y **c)** diagramas de dispersión con suavizadores en la dirección *y* y *x* en el triángulo inferior, resaltando altas correlaciones por líneas casi paralelas.

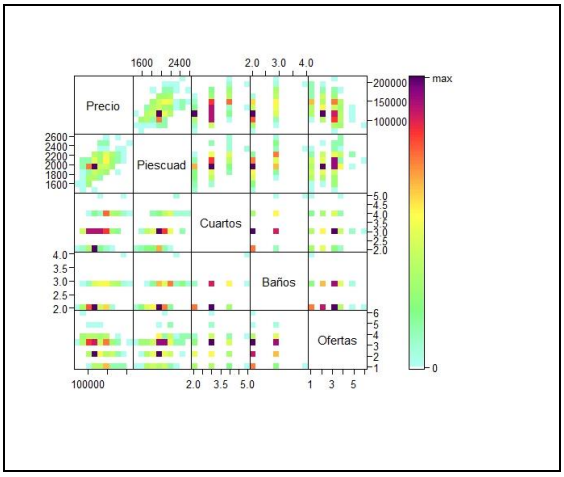




La sentencia **iplot** produce una imagen del diagrama de dispersión de grandes volúmenes de datos donde los colores codifican la densidad de los puntos en el diagrama de dispersión. También funciona con factores.

La sentencia **ipairs** produce una imagen de una matriz de diagramas de dispersión para grandes volúmenes de datos donde los colores codifican la densidad de los puntos en los diagramas de dispersión.



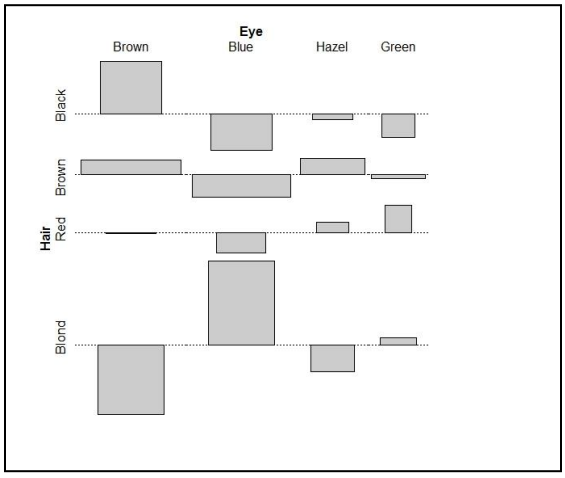


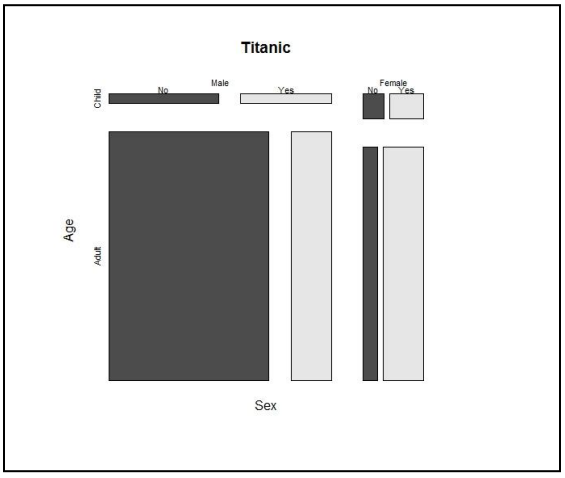
Autor: Dr. Erwin Kraenau - Lic. José Cárdenas

El **diagrama de mosaico** es un método gráfico para la visualización de los datos de dos o más variables cualitativas. Es la extensión multidimensional de spineplots (se utiliza cuando un conjunto de datos proporciona una variable dependiente nominal y una variable de intervalo), que gráficamente muestra la misma información para una sola variable. Se da una visión general de los datos y hacen posible identificar las relaciones entre las diferentes variables. Por ejemplo, la independencia se muestra cuando las cajas a través de categorías tienen las mismas áreas. Los diagramas mosaico fueron introducidos por Hartigan y Kleiner en 1981 y ampliado por Friendly en 1994.

Como con gráficos de barras y spineplots, el área de los rectángulos, también conocido como el tamaño del bin, es proporcional al número de observaciones dentro de esa categoría.

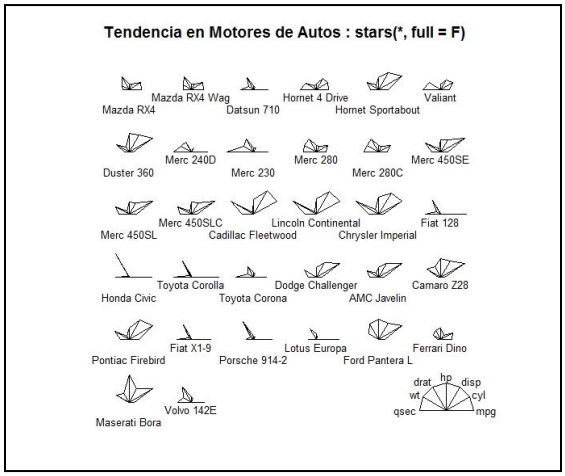
Las sentencias **assoc** y **mosaicplot** producen diagramas de asociación indicando las desviaciones de un modelo de independencia especificado posiblemente en una tabla de contingencia de alta dimensión.

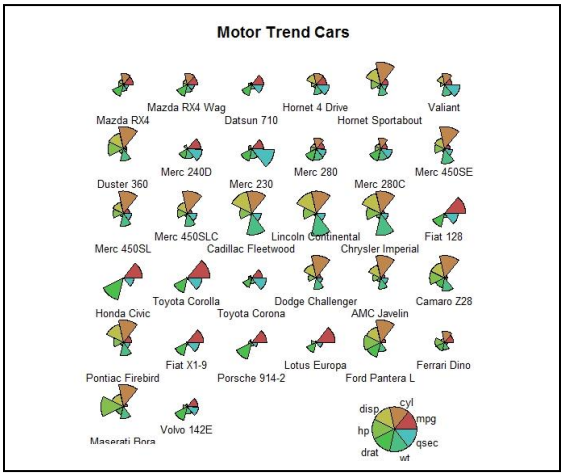




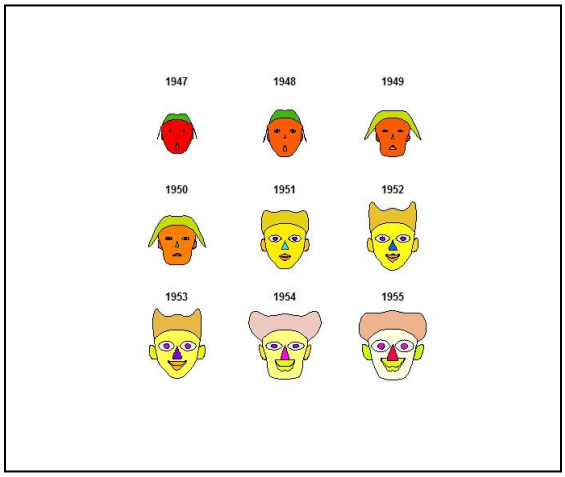
Supóngase un conjunto de datos multivariados ordenados matricialmente, de manera que las filas corresponden a las observaciones y p columnas, una por cada variable. En dos dimensiones se pueden construir círculos (uno por cada observación multivariada) de un radio prefijado, con *p* rayos igualmente espaciados emanando del centro de cada círculo. Las longitudes de los rayos son proporcionales a los valores de las variables en cada observación.
Los extremos de los rayos pueden conectarse con segmentos de líneas rectas para formar una estrella. Con cada observación representada por una estrella, éstas pueden ser agrupadas según sus similitudes. Es conveniente estandarizar las observaciones, caso en el cual pueden resultar valores negativos.
(Johnson, Wichern (1998)).

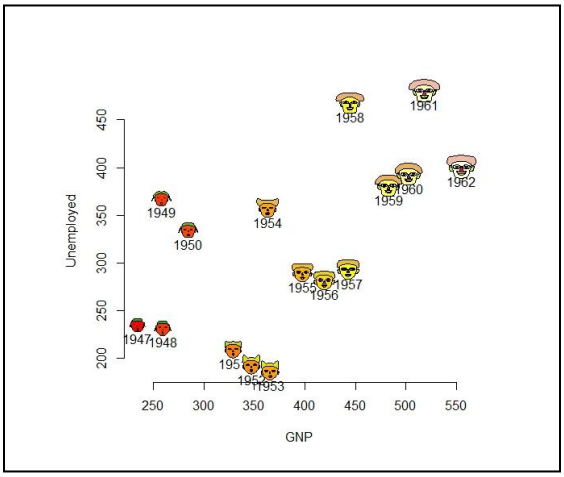
Para los gráficos de estrellas se utiliza la sentencia **star**.





Las **caras de Chernoff** Gráfica que muestra datos multivariados en forma de caras humanas. Los rasgos característicos individuales como los ojos, la boca, la nariz representan valores de las variables a través de su forma, tamaño, lugar en la cara donde se encuentran ubicados y orientación. La idea que sustenta este tipo de representación gráfica es la facilidad con que el observador reconocerá cualquier variación por imperceptible que parezca. Ya que los rasgos de la cara cambian muy considerablemente, se debe escoger cuidadosamente la forma en que se grafican los rasgos.
Las caras de Chernoff se obtienen mediante la sentencia **faces**.



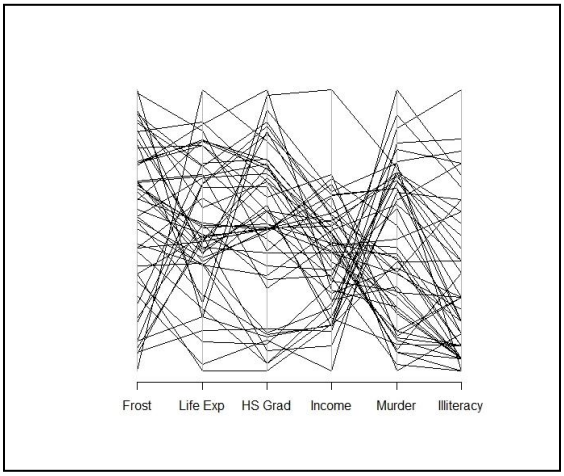


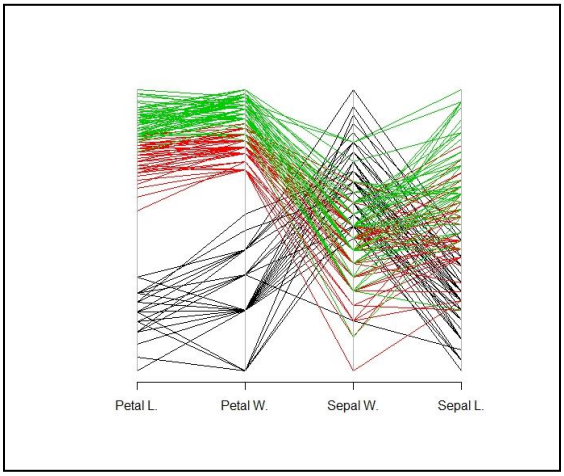
Las **coordenadas paralelas** es una manera muy común de visualizar la geometría de alta dimensión y analizar datos multivariantes.

Para mostrar un conjunto de puntos en un espacio *n*-dimensional, un fondo es trazado que consiste de *n* líneas paralelas, típicamente vertical e igualmente espaciados. Un punto en el espacio *n*-dimensional es representado como una poli

línea con vértices sobre los ejes paralelos; la posición del vértice sobre el eje *i*-ésimo correspondiente a la coordenada *i*-ésima del punto.

La visualización está estrechamente relacionada a la visualización de las series de tiempo, excepto que es aplicado a los datos donde los ejes no corresponden a puntos en el tiempo, y por tanto no tienen un orden natural. Por lo que, diferente arreglos de los ejes pueden ser de interés.





La sentencia **plotsummary** muestra algunas características importantes de las variables de un conjunto de datos. Para cada variable un diagrama es calculado que consiste de un diagrama de barras, la función acumulada, traza una densidad y un diagrama de caja.