Visualizing Statistics

Pedro Valero Mora-valerop@uv.es

Metodología de las CC del Comp-Universitat de València

Abril 2010

VNIVERSITAT DE VALÈNCIA

Indice

Introducción y Ejemplos

Introducción

"Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught."

John W. Tukey, We need both exploratory and confirmatory,

The American Statistician, 34(1), (Feb., 1980), pp. 23-25.

- No obstante, cuando utilizamos software estadístico estándar a menudo nos encontramos con un conjunto de técnicas y métodos ya prediseñados
  - Hay poca flexibilidad
  - No están basados en gráficos

- Determinados sistemas estadísticos no obstante permiten un buen grado de flexibilidad: Análisis de datos interactivos
  - R permite un grado de flexibilidad que otros sistemas a menudo no tienen
- ¿Y cuándo se trata de gráficos?
  - El desafio es que los gráficos se conviertan en el análisis, no en la parte final de éste
- Instalemos ViSta y veamos un ejemplo

Instalación de ViSta

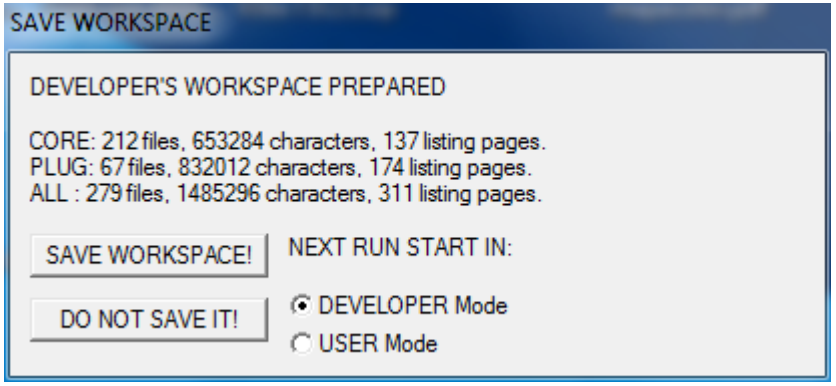
1. Descargar ViSta de http://www.uv.es/visualstats/Book/ e ir a la sección de descargas
2. Descargar en

   www.uv.es/visualstats/Book

   (el lugar original www.visualstats.org y similares no están actualizados)
3. Descomprimir en donde se quiera utilizar (es un archivo .zip)
4. Poner en marcha ViSta.exe

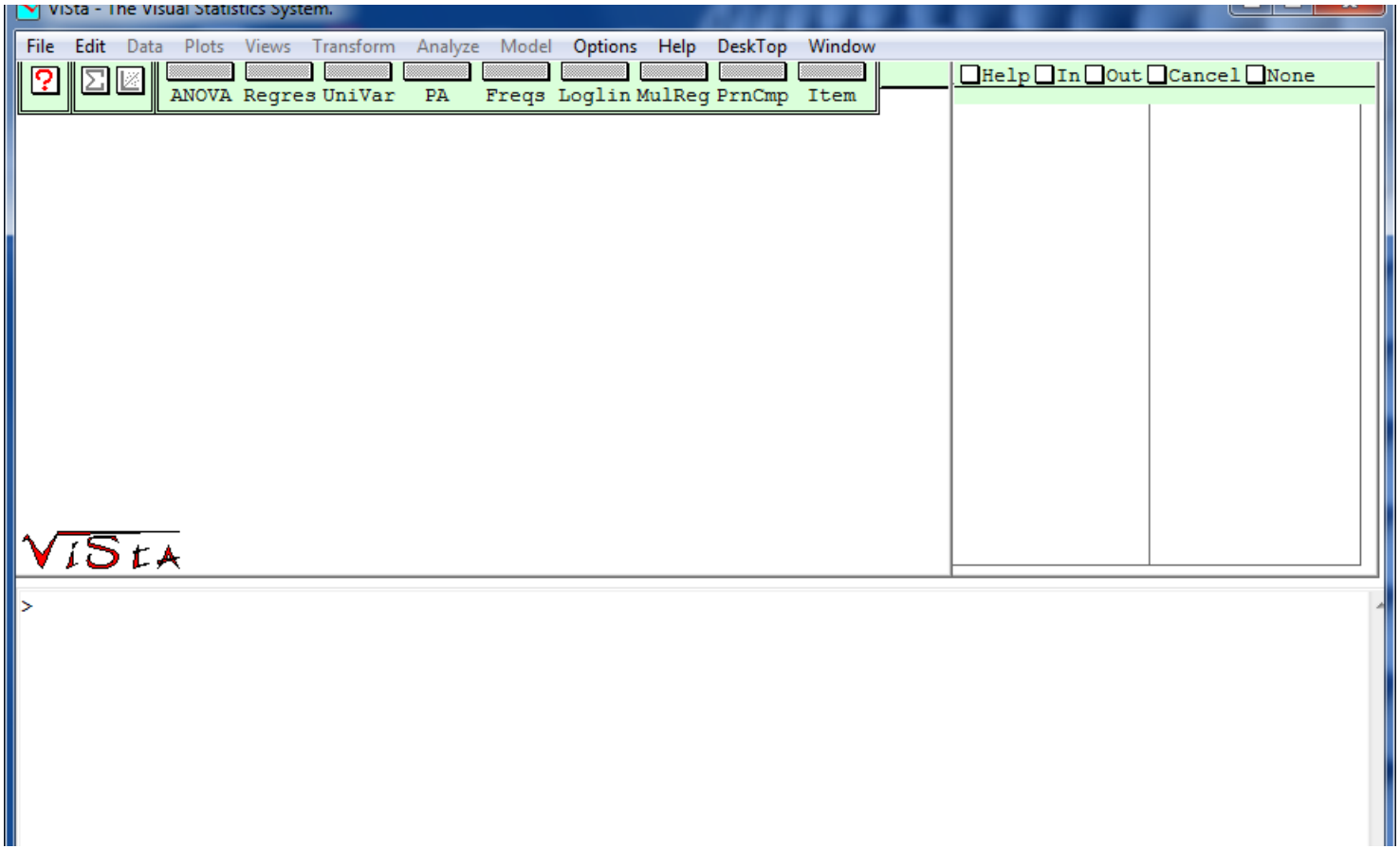
   Se pondrá en marcha el proceso de carga de archivos y directorios

   Aparecerá el siguiente cuadro de diálogo.



5. Hacer click en SAVE WORKSPACE

6. Volver a poner en marcha ViSta.exe



7. Ir al menú File>Open Data

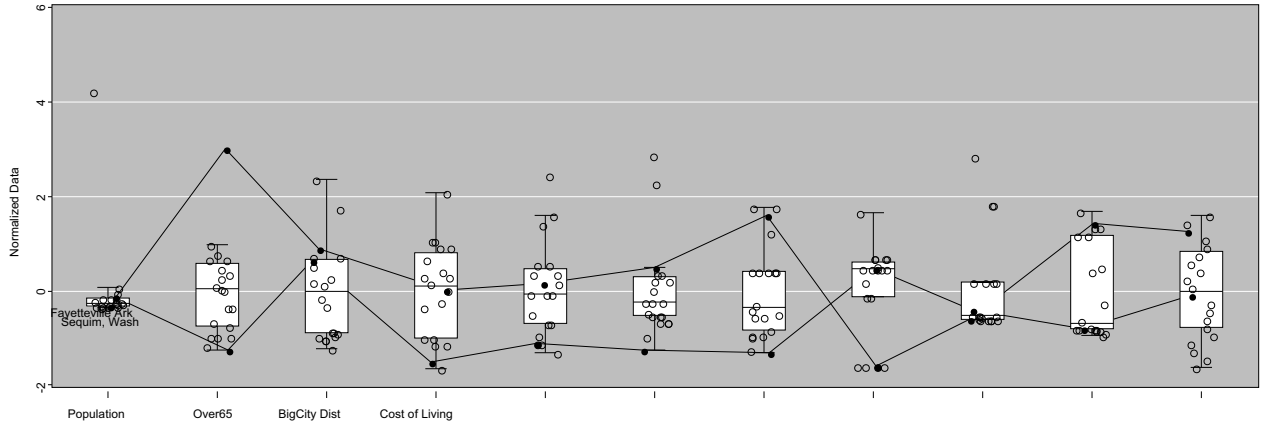
   En el cuadro de diálogo hacer data>regress>retire.vdf

   Esto nos dará unos datos para empezar

Ejemplo: Ciudades para jubilación

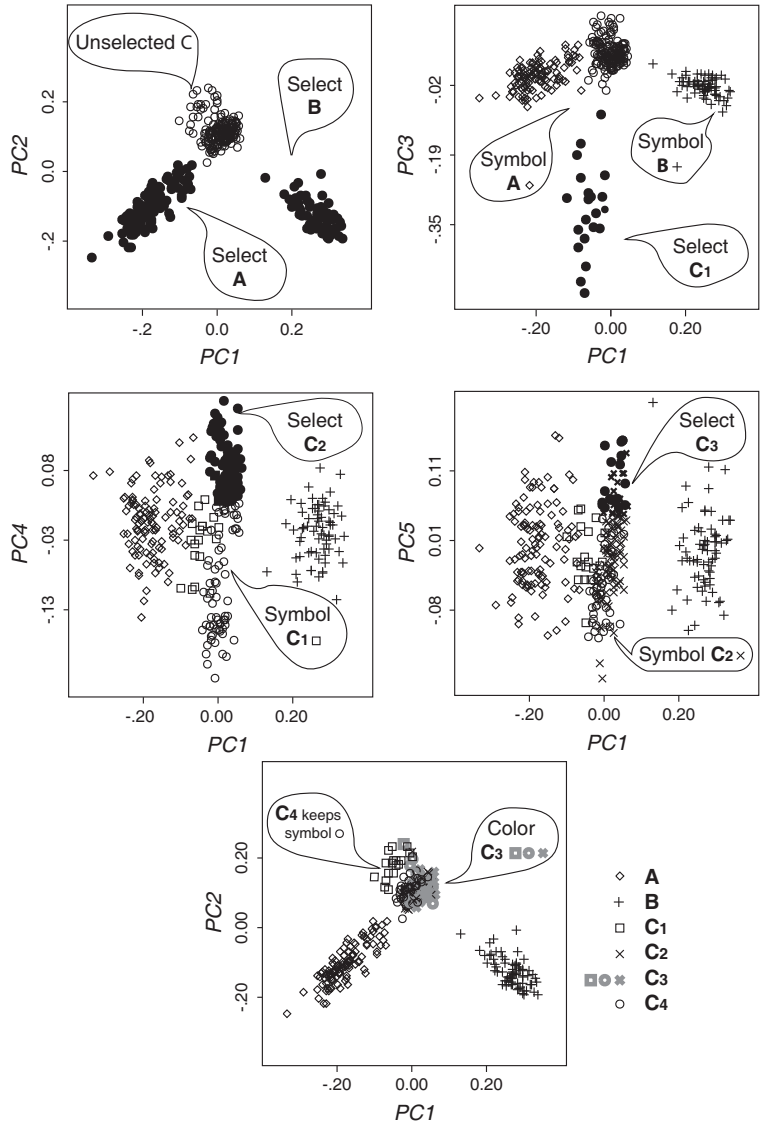
- Seleccionar **Boxplots** en el menú **Plot**

  Obtendremos un gráfico de este tipo



- Preguntas que podemos explorar
  - ¿En qué difieren la ciudad más jóven y la más anciana?
  - ¿Cuál es la mejor ciudad globalmente?

Ejemplo: Identificando clusters visualmente



Historia y Software

Historia

- La idea de los gráficos dinámicos puede decirse que empezó con PRIM9

  No obstante, se utilizaba hardware especial sólo disponible en ciertos lugares

- Buena parte de los desarrollos iniciales acabaron en el libro "Dynamic Graphics for Statistics"

- Hasta la aparición de los ordenadores Macintosh, no fue posible llevar este tipo de análisis a todo el mundo. Un par de aplicaciones que marcaron el camino a seguir fueron:

  MacSpin

  Statview

Software (más o menos actual)

Comerciales:

- DataDesk
- JMP
- Tableau
- Spotfire

No comerciales

- ViSta
- Arc
- XGobi, GGobi y RGobi
- Manet y Mondrian
- iPlots

En la nube

- Google Chart Explorer (no muy interactivo)
- Gapminder
- Google public data explorer
- TrendCompass

LispStat

- LispStat fue un lenguaje de programación estadístico (similar a R) desarrollado en los años 90 por Luke Tierney.

  ViSta está desarrollado practicamente del todo en LispStat

- En su momento ofrecía la posibilidad de experimentar con gráficos estadísticos interactivos y dinámicos a un nivel que no había sido posible previamente

  Algunos creen que esa facilidad no ha sido igualada todavía

- La gente que desarrollaba software para Lisp-Stat se centraba a menudo en gráficos sofisticados, dinámicos

  Los resultados de texto parecían secundarios

- Lisp-Stat está actualmente moribundo pero parece que hay algunos esfuerzos por revivirlo

  Lisp-Stat: Past, Present and Future

  Back to the Future

  CommonLispStat

  Incanter

A modo de resumen

- La cantidad de recursos disponibles es muy grande actualmente
- No obstante, las ideas básicas son las mismas practicamente desde hace 20 ó 30 años
  - Aunque esas ideas aplicadas tienen muchas posibilidades
- Software comercial versus software no comercial
  - El comercial ofrece técnicas más probadas, más simples pero no necesariamente las mejores
  - El no comercial ofrece técnicas más avanzadas pero el esfuerzo necesario para utilizarlo es a menudo mayor (aunque R parece ser capaz de cambiar el escenario de una manera radical)


ViSta

¿Por qué ViSta?

- ViSta incorpora buena parte de las técnicas interactivas que se han propuesto desde los años 80
- El código es abierto y realizar modificaciones es sencillo
- Integra técnicas estadísticas y gráficos en un mismo entorno

  Varios de los sistemas que hemos visto necesitan dos lenguages: Por ejemplo, R y C o R y Java,
- Es una manera sencilla de familiarizarse con las técnicas
  - ViSta no es perfecto pero tiene muchas cosas interesantes
  - No obstante, hay que tener en cuenta que este curso está centrado en los *conceptos:* No tenemos la capacidad de una empresa comercial

Historia

- Desarrollado por Forrest W. Young
  - 15 años de desarrollo! Desde el año 1991
  - Fundamentalmente usando XLispStat
  - Muchas características, algunas de ellas posteriormente eliminadas
- Desde el año 1998 me incorporo al proyecto haciendo cosas de
  - Datos "missing"
  - Modelos loglineales
  - Regresión múltiple (sin terminar)
  - Adaptaciones de MDS, Cluster, Mapas
  - Multitud de corrección de bugs, detalles, etc.

- Rubén Ledesma también ha incorporado un buen número de características

  Adaptación del módulo de análisis de homogeneidad

  Software para Psicometría

  Bootstrap

  etc.

- En 2006 publicamos el libro "Visual Statistics: Seeing your data with dynamic interactive graphics"

- Michael Friendly incluye ViSta y LispStat en su selección de importantes momentos en la historia de los gráficos estadísticos

Características

- Gratuito
- Gráficos múltiples, interactivos, dinámicos
- Funciona en Windows (versiones anteriores funcionaban en Mac y Unix)
- Workmap (una representación de los pasos realizados)
- Spreadplots (varios gráficos simultaneamente)
- GuideMaps (desconectado en la versión actual)
- Hoja de datos (muy sencilla pero permite ver los datos)
- Tipo de datos (lo veremos ampliado en la siguiente sección)
- Ayuda

- Ampliable
- Gráficos en formato vectorial
- ¿Ya he dicho gratuito?
- El proyecto está abierto a cualquiera con interés en desarrollar nuevas cosas, así como cualquier otra tarea!!

  Si teneis datos que podamos analizar, estamos interesados!!

Tipo de datos

- Una característica de ViSta es que atribuye un tipo a los conjuntos de datos a partir de las características de las variables que hay en él:
  - Sólo variables numéricas: Numérico

    Si tiene algún valor perdido: Missing (se usa nil)
  - Una variable numérica y una o varias categóricas: Clasificación
  - Sólo variables categóricas

    Sin agrupar: Datos categóricos

    Agrupadas y con una variable denominada Freq o Frequency: Frequency classification

    Cruce de dos variables categóricas: Frequency Table
  - Datos de similaridades/disimilaridades: Datos relacionales

- ViSta limita los análisis a los tipos de datos y lo hace sobre la marcha (a partir de lo seleccionado).

  A menudo lleva a que los gráficos/análisis se seleccionen automáticamente

  A veces, no obstante, producen más problemas que los que solucionan!

Interacción con gráficos

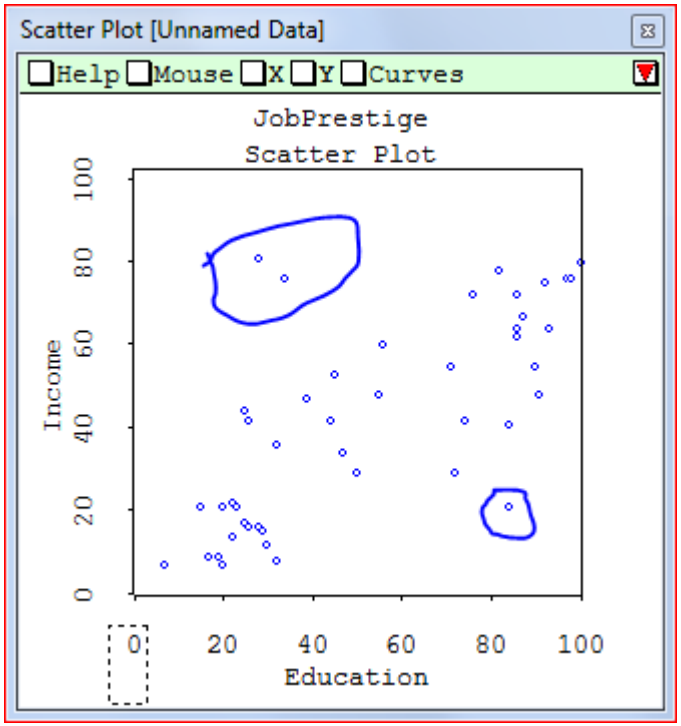
Ejemplo: Jobs

- Este ejemplo está en data/regress/jobs.vdf

  Las variables de Income, Prestige y Salary están escaladas a 100

Activación de elementos

- En el gráfico de Income versus Education podemos ver tres puntos que no parecen ir con el resto



- Cepillado: Pasar el ratón para ver los nombres
- Selección: Cambiar el modo del ratón para ver los nombres
- Selección múltiple: Utilizando Ctrl seleccionar varios puntos
- Selección de area: Arrastrar

Etiquetas

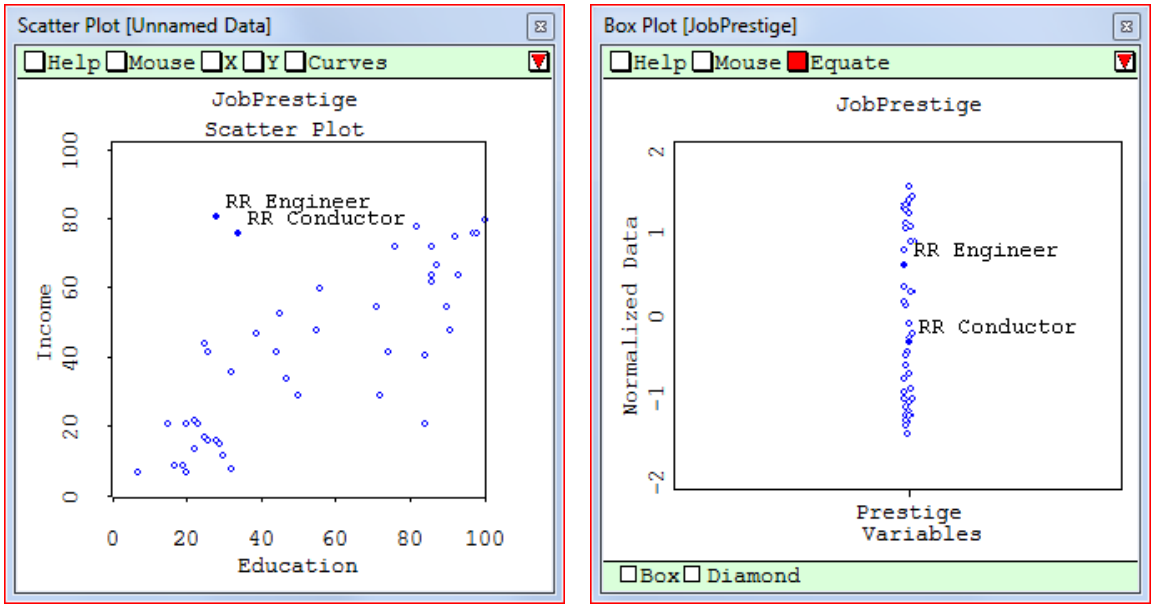
- Etiquetar los casos es una de las necesidades más básicas en muchos análisis
  - No obstante, poner las etiquetas automáticamente es muy difícil
  - Por ello, que las etiquetas se muestren interactivamente es de gran utilidad
  - Esto es sobre todo importante en gráficos densos, en los que se necesita explorar partes en detalle
- En ViSta, al seleccionar se muestra la etiqueta del punto

  Se puede desconectar la opción de mostrar etiquetas y verlas en la lista de etiquetas del archivo de datos

Ligado

El ligado es una de las estrategias más potentes. Las acciones sobre un gráfico se propagan a los otros gráficos

En el ejemplo de Jobs podemos conectar el diagrama de dispersión a un diagrama de puntos



Enfocar y Excluir

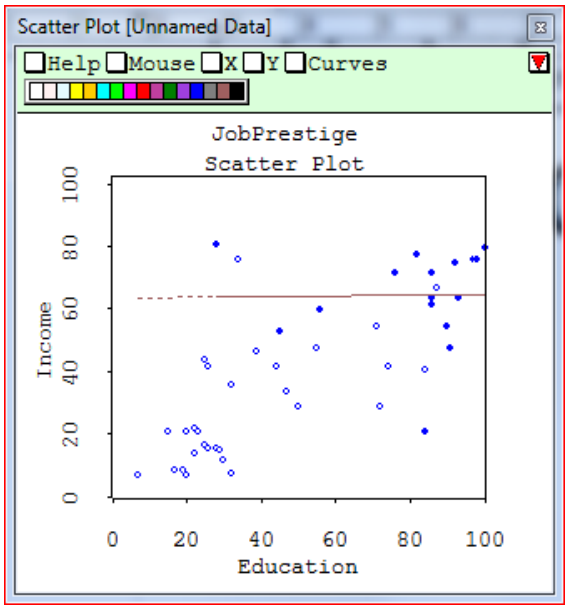
- En el ejemplo de Jobs hay un par de puntos que parecen no ir con el resto.
- En el menú contextual se pueden excluir esos puntos
  - **Remove Selection** excluye los puntos seleccionados.
    Las escalas se ajustan automáticamente a los puntos que quedan
  - **Focus** se centra en los puntos seleccionados
  - **Show all** muestra todos los puntos de nuevo
- Esas características están ligadas así que los otros gráficos ven reflejados los cambios

Cambiar Colores y Símbolos

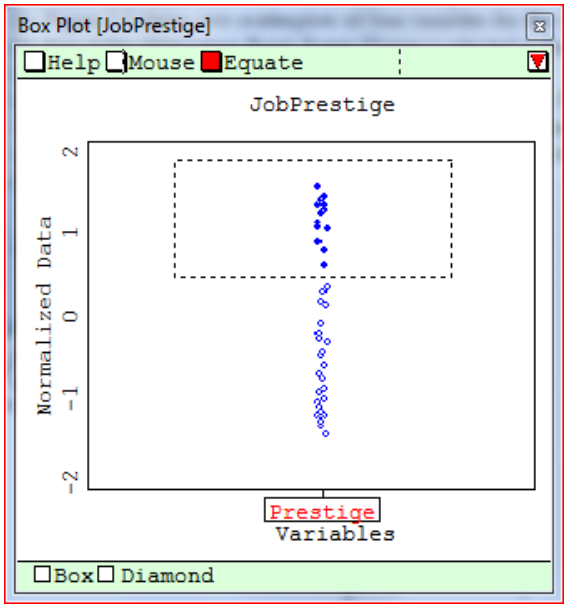
- Usando **Color Pallet** y **Symbol Pallet** aparece una paleta de colores
  - Haciendo click sobre el color, los puntos seleccionados cambian de color
- Esas características están ligadas así que se ven reflejados los cambios de un gráfico en los otros

Ejemplo: Cepillado, ligado y líneas

- En el gráfico elegir Curves y seleccionar Regression Lines selected



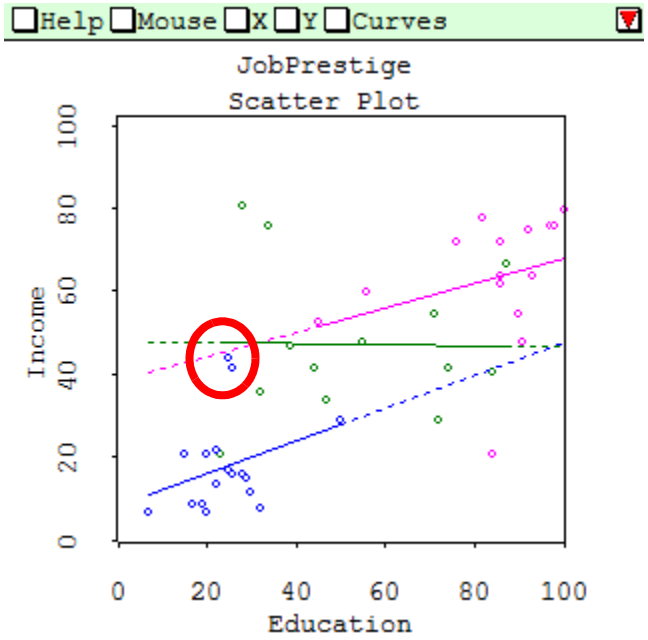
- En el gráfico de puntos, hacer el cepillo más grande y cepillar de arriba abajo



La línea de regresión irá cambiando, ajustándose a los puntos seleccionados

- ¿Qué podemos aprender de este ejercicio?
  - Para los niveles medios de prestigio, no parece que la Educación tenga efectos sobre los ingresos
  - Podemos hacer tres grupos con el Prestigio y cambiar el color
  - Luego, en Curves, podemos pedir Regresión por color

- ¿Y los valores extremos?
  - Podemos probar a quitarlos y ver que pasa
  - Cuando quitamos dos puntos influyentes en los valores medios de prestigio vemos que la relación entre educación en ingresos es la misma para todos los grupos de prestigio
  - Al hacerlo descubrimos nuevos valores destacados en el grupo con baja educación: Plumber y Tram Motorman



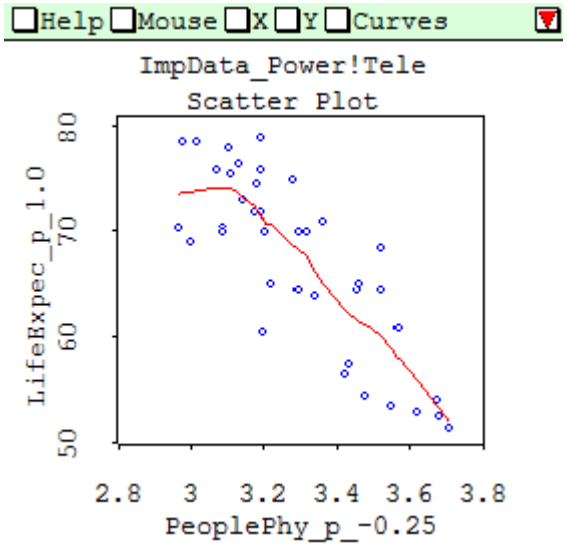
Ejemplo: Transformación, Imputación y Visualización

1. Abrir el ejemplo Tele.vdf en data/regress
2. Usar las transformaciones BoxCox para linearizar

   ¿Corea del Norte?
3. Imputar los valores perdidos
4. Visualizar la relación entre número de médicos per capita y esperanza de vida

   Lowess aporta algo interesante

   Usando un gráfico de puntos para número de televisiones per capita puede explorarse la influencia de la economía



Gráficos básicos

Gráficos de presentación v. gráficos de análisis

- Gráficos de presentación

  Los gráficos de presentación intentan dar una idea final completa de un resultado

  Son una ilustración

  A menudo empleamos bastante tiempo en ajustar cada detalle

  No es posible ir más allá o quizás sí?

  - El considerado mejor ejemplo de gráfico es el de Minard

    Pero se puede ver que incluso el mejor de todos podría ser mejorado

- Gráficos de análisis

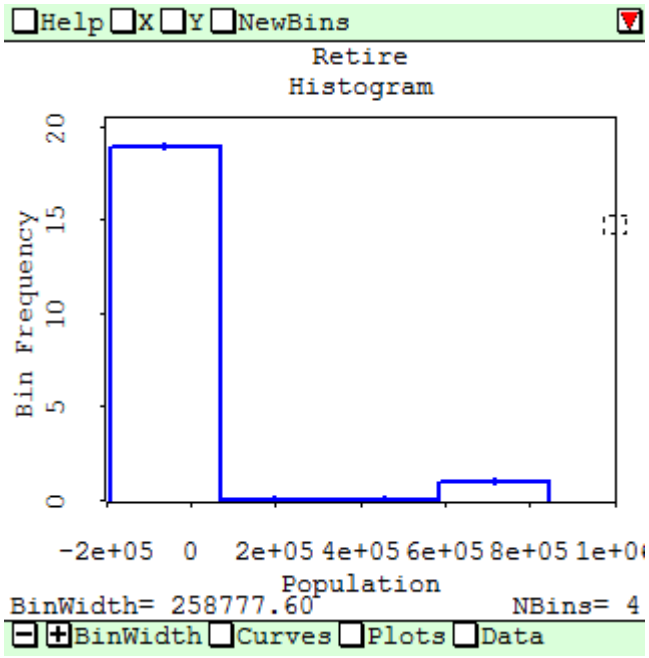
  Permiten ir más allá de lo representado

  Jugar con aspectos importantes

  Añadir elementos, quitarlos. El artículo de Weisberg da ejemplos.

Histogramas

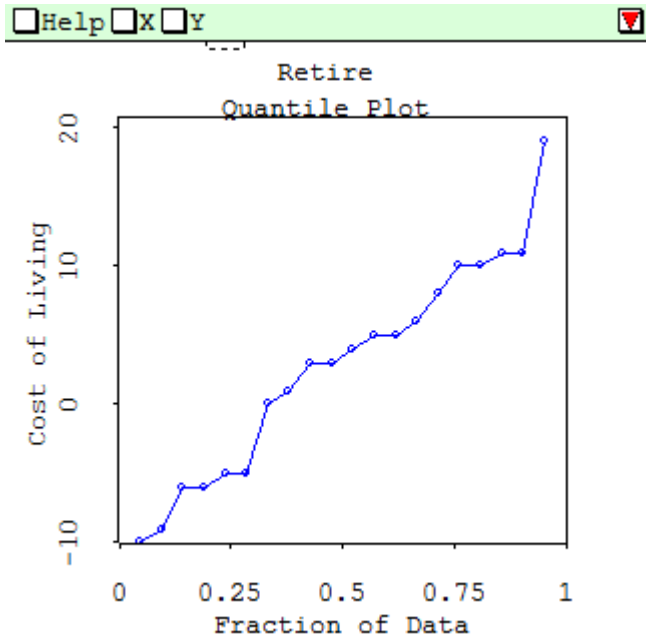
- Cuatro comandos (pero tres son iguales!)



- Botón X, cambiar de variable
- Botón Y, pasar a probabilidades
- Binwidth: permite cambiar y explorar el efecto que tiene cambiar el tamaño de las barras
- Curves: Añade curvas de estimación de la densidad, controladas por una barra de desplazamiento
- Plots: Cambia de tipo de gráfico

Plot acumulativo

- Gráficos de cuantiles



Gráfica los valores de la variable frente al cuantil

Para datos simétricos, la línea central se aproximará a la diagonal

Para datos asimétricos positivos, la línea irá por debajo de la diagonal

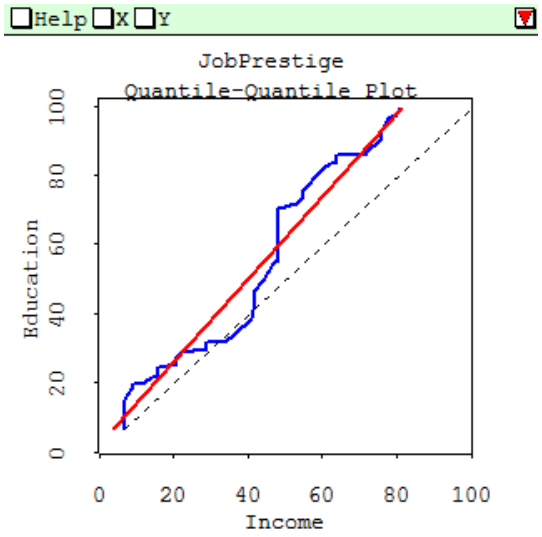
Lo contrario para datos positivos negativos

- Haciendo click en X cambia a un gráfico de probabilidad normal

Si la distribución es normal, los valores seguirán la diagonal

Comparación de dos variables

- Este gráfico permite comparar las distribuciones de dos variables



- El gráfico representa los cuantiles de una variable frente a los cuantiles de la otra variable

- La línea azul indica si las dos variables tienen distribuciones con la misma forma
- La línea roja representa dos variables dos variables con la misma forma pero con centro y amplitud iguales a las variables observadas
- La línea de puntos representa dos variables semejantes

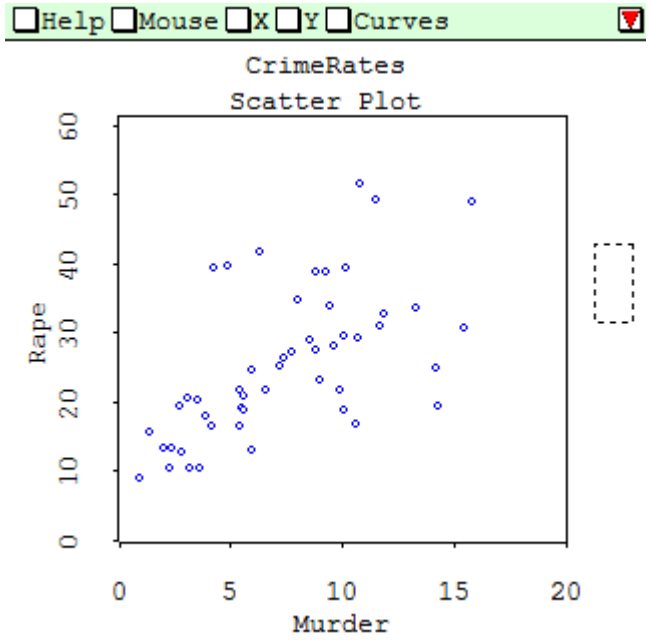
- Interpretación
  - Línea azul igual a línea de punto: igual forma, centro y amplitud
  - Linea azul igual a roja pero no a línea de puntos: igual forma

Gráfico de Puntos

- Muy similar al gráfico de líneas (creo que lo desconectaré en el futuro)

Diagrama de dispersión

- Los botones X e Y sirven para cambiar las variables que se incluyen
- El botón Curves sirve para añadir una variedad de líneas



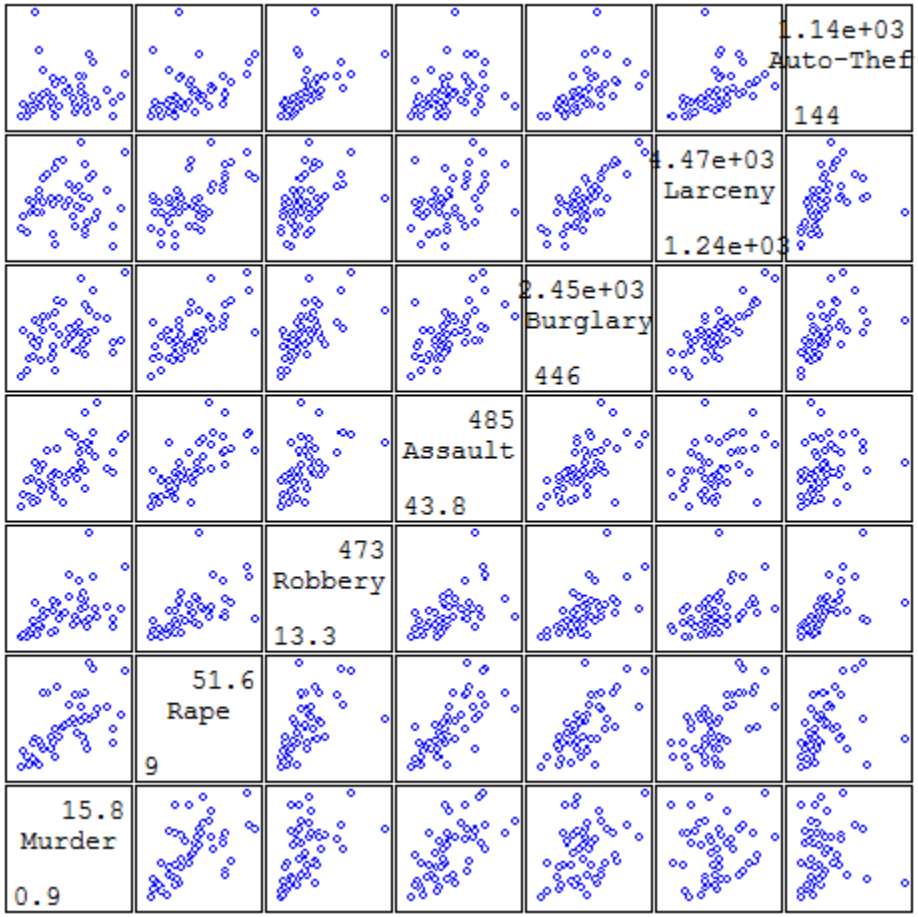
- Las líneas disponibles son:
  - Eje Principal
  - Líneas de regresión
  - Línea monotona
  - Normal and Quantile contours
  - Lowess: regresión local

    El slider controla la proporción de casos que es utilizada
  - Kernel smoother: otra forma de regresión local

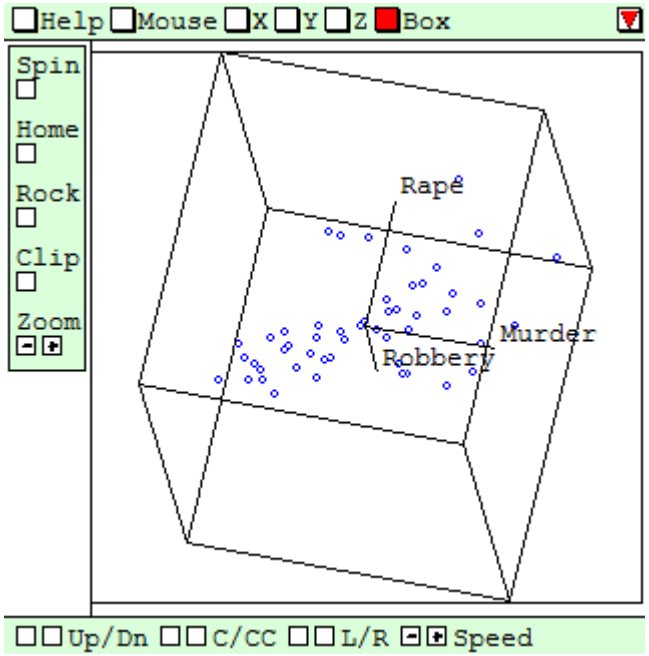
    El slider controla la anchura de la función kernel

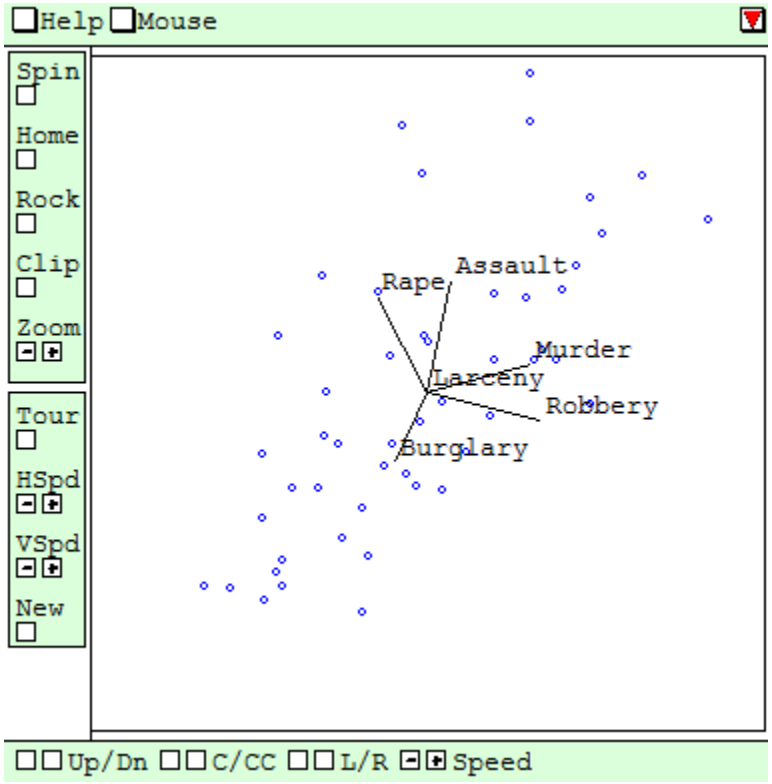
Matrices de diagramas de dispersión

- Muestra los diagramas de dispersión para varias variables simultaneamente



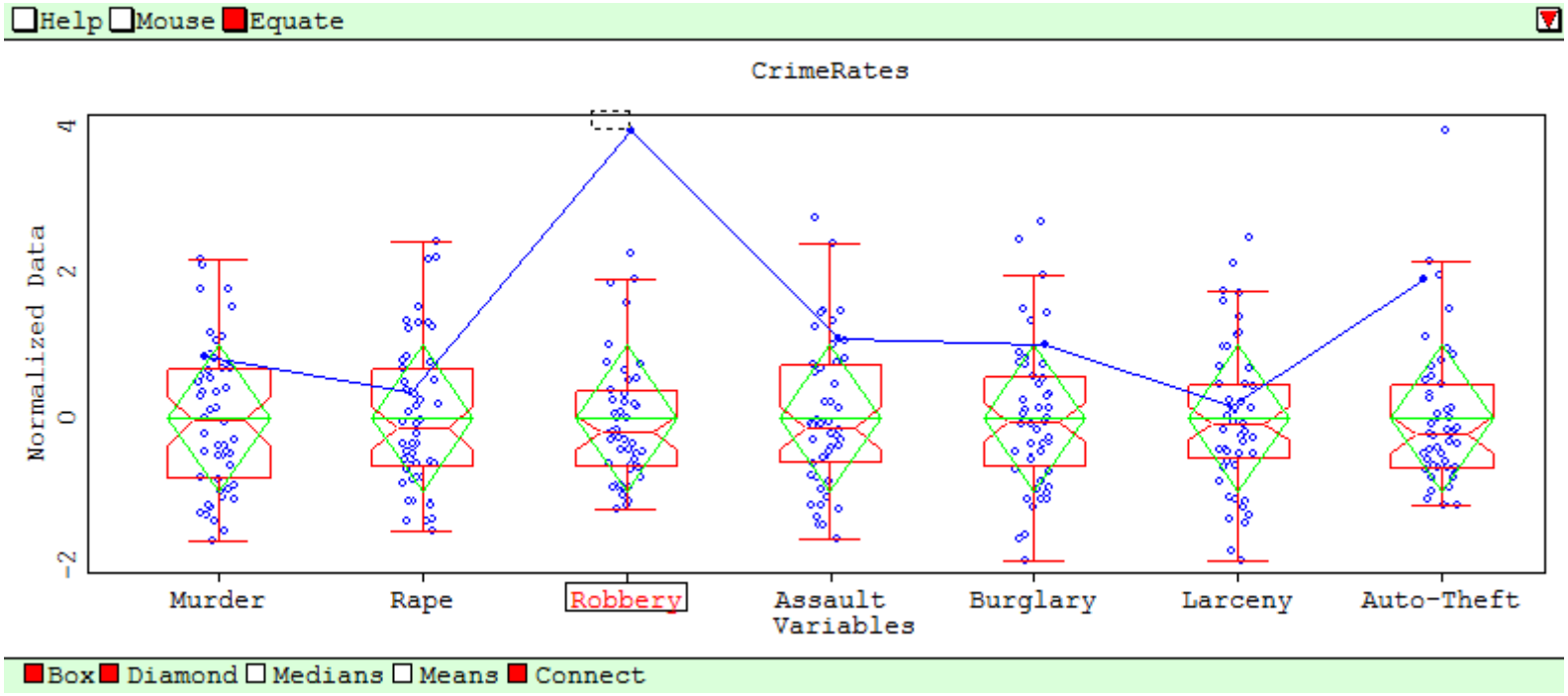
Spinplots y Tourplots





Boxplot, Diamond plot, Parallel coordinates

- Son variantes del mismo gráfico

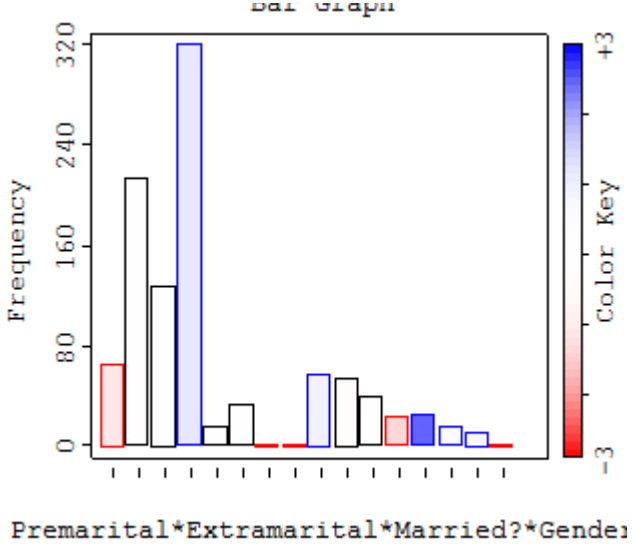


- Los diamantes son las medias. Los casos están conectados.

  El botón Equate estandariza o no las variables. Se pueden añadir líneas para medianas, medias.

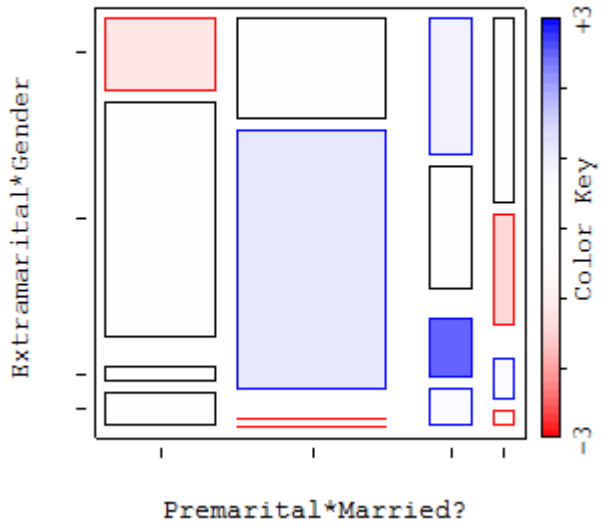
Gráficos de mosaico y diagramas de barras

- Son gráficos para datos categóricos



No están ligados (tendría que hacerlo esto)

No tienen la posibilidad de cambiar de variables (por eso es mejor utilizarlos desde la visualización para datos de frecuencias)



Spreadplots

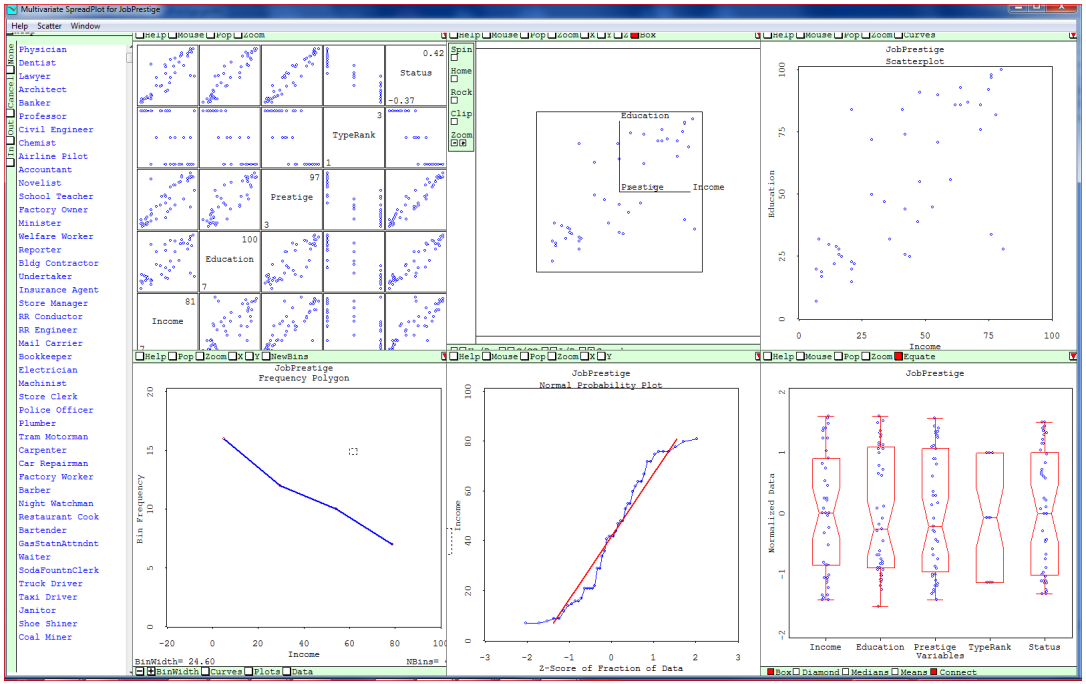
Manejando muchas ventanas

- A veces tenemos ideas que usan muchas ventanas/gráficos

  Además, la interacción entre ellas puede ser sofisticada

  Si tenemos una combinación de ventanas interesante podemos querer guardarla para no tener que repetir el proceso cada vez

- Una solución son los spreadplots

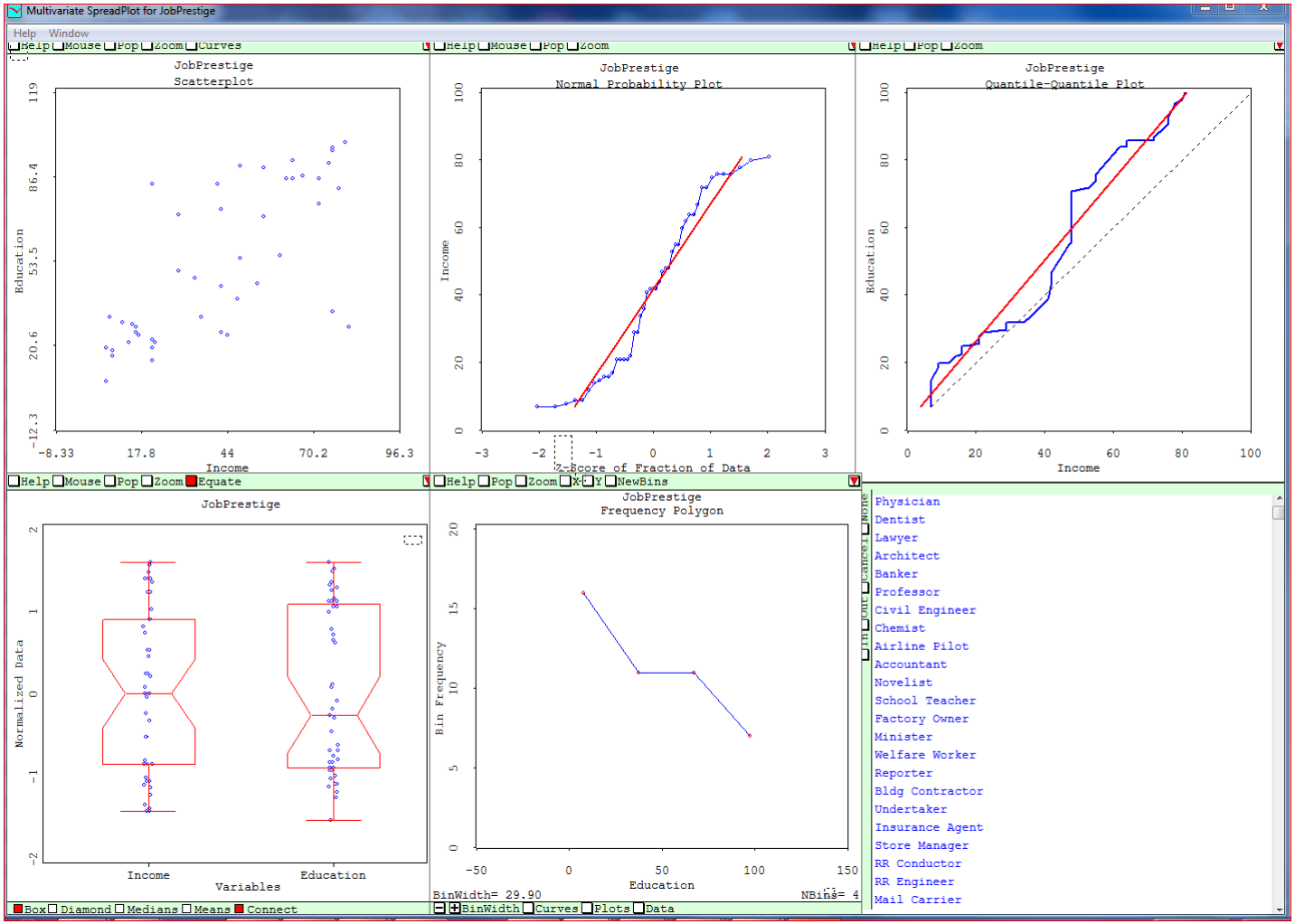


Producen una combinación de gráficos ajustada a un problema concreto

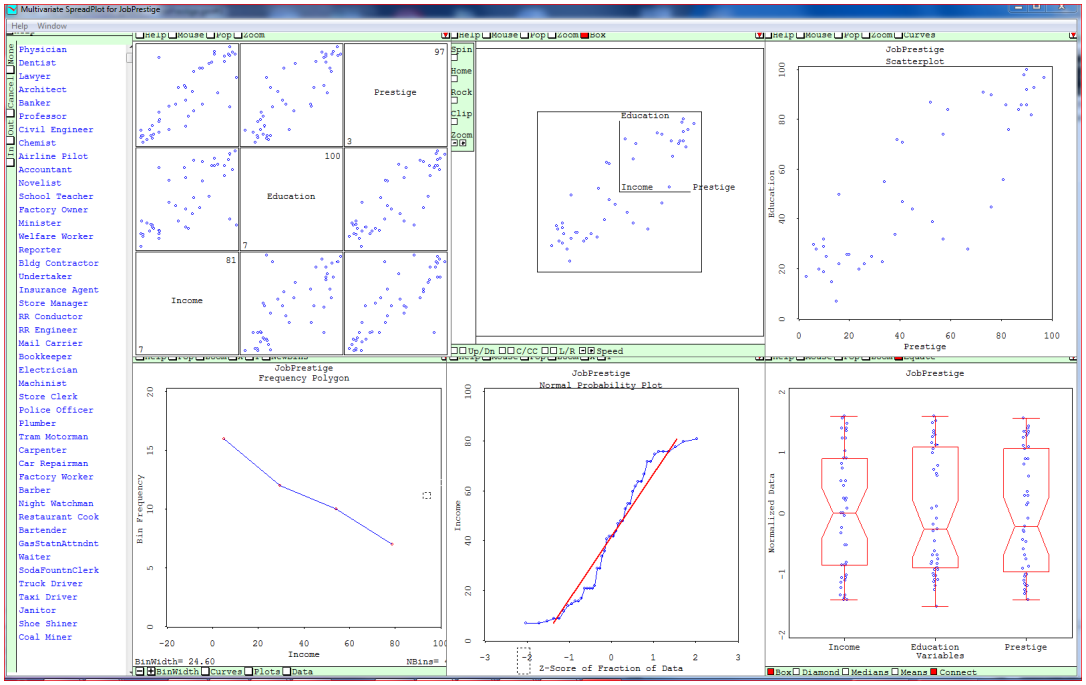
Las ventanas funcionan conjuntamente (se abren, se cierran, etc.)

Son programables

Spreadplot para 2 variables numéricas

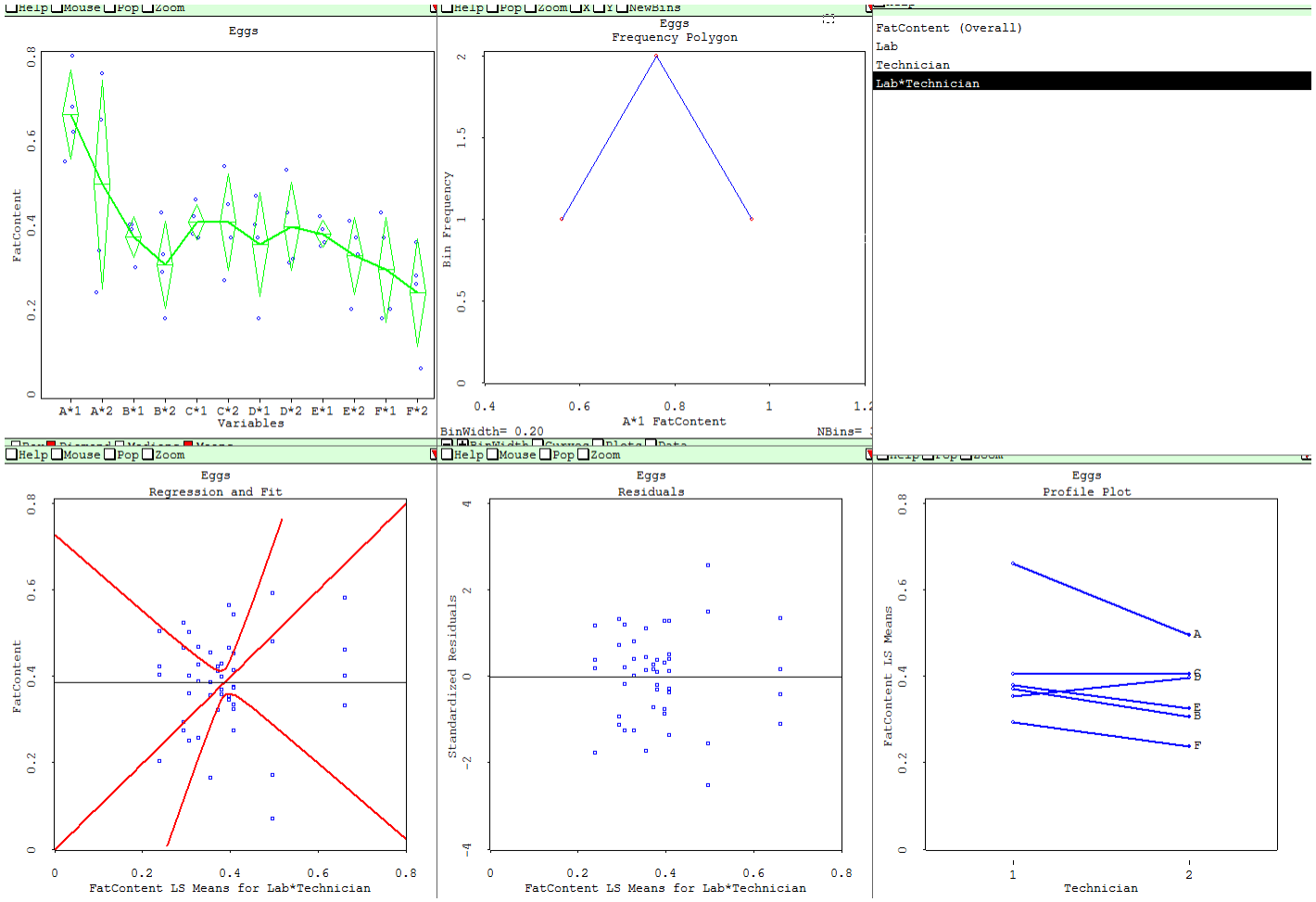


Spreadplot para 3 variables numéricas

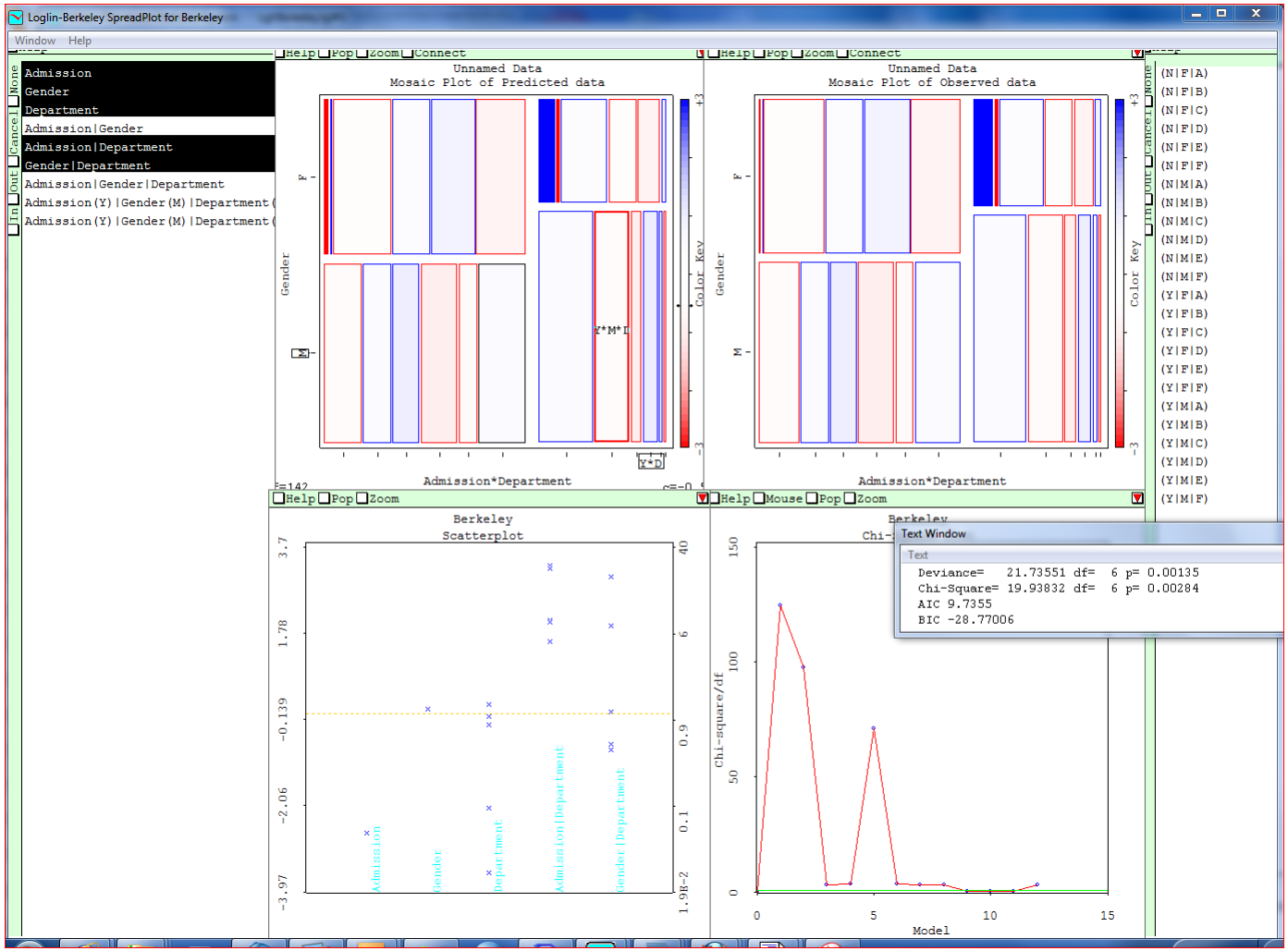


Fijarse que la matriz de diagramas de dispersión sirve de control para el resto de gráficos

Analisis de varianza



Spreadplot para modelos loglineales



Notas finales

- La idea de los spreadplots es muy poderosa

  En JMP y en DataDesk hay conceptos muy parecidos

  Varios programas comerciales también los utilizan (Spotfire, TrendCompass)

  Dashboards es un concepto muy similar que parece estar de moda

Datos categóricos

Visualización de datos categóricos multivariados

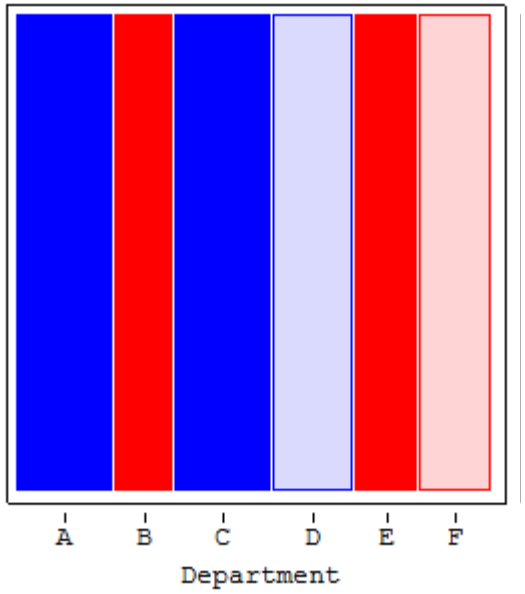
- El desarrollo de las técnicas de visualización se han basado sobre todo en datos cuantitativos
  - Los datos categóricos han tenido un desarrollo más lento: ver Friendly, 2000
  - Las extensiones dinámicas han sido exploradas todavía más recientemente

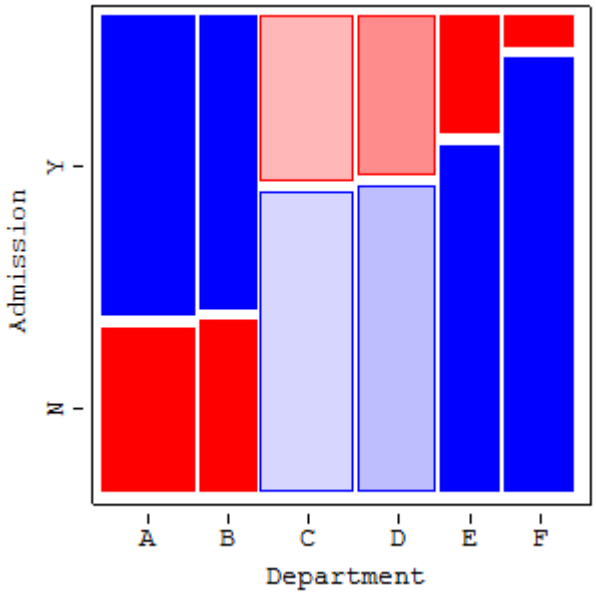
    Manet fue el primer programa en explorar este tema sistemáticamente
- Los plots de mosaico han generado mucho interés por su capacidad de representar muchas variables categóricas simultáneamente

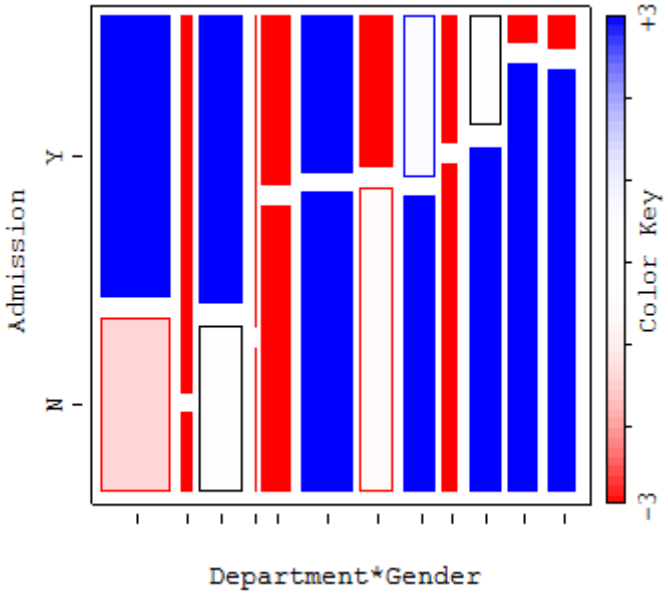
  En principio, el número de variables que se pueden incluir es ilimitado

  No obstante, ViSta está limitado a cuatro

- Un ejemplo de plot de mosaico







- Se trata de los mismos datos con tres variables

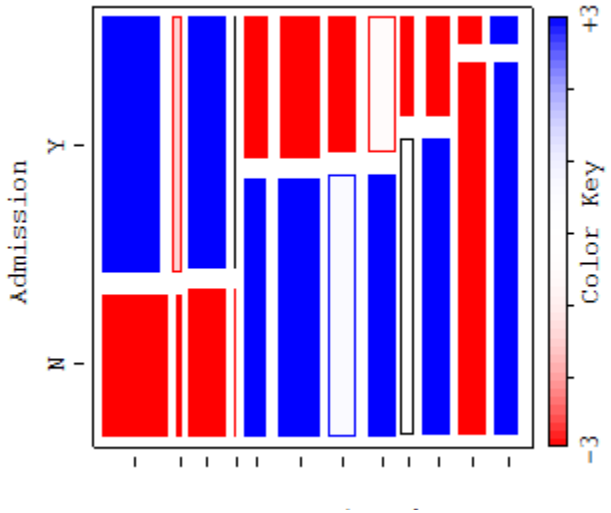
  Cada vez que se añade una variable, el gráfico se subdivide

  Cada celda es proporcional a la frecuencia condicional de las categorías a las que corresponde

  El color de las celdas es el residual standarizado respecto de un modelo. En este caso se trata del modelo de efectos simples (azul +, rojo -)

Spinogramas y Mosaic plots

- Originalmente ViSta hacía Mosaic plots, pero lo he cambiado (trabajo en progreso)



La diferencia sólo se nota cuando hay más de dos variables

- Los Spinogramas son más apropiados cuando tienes una variable dependiente y la situas en el eje Y

Es decir son más apropiados para modelos logit

- Los plots de Mosaico son más acordes con la situación en la que no se distingue entre variables dependientes/independientes

  Más apropiados para modelos loglineales

Ejemplo: Datos de Berkeley

- Los datos de Berkeley analizan la discriminación por género por departamento

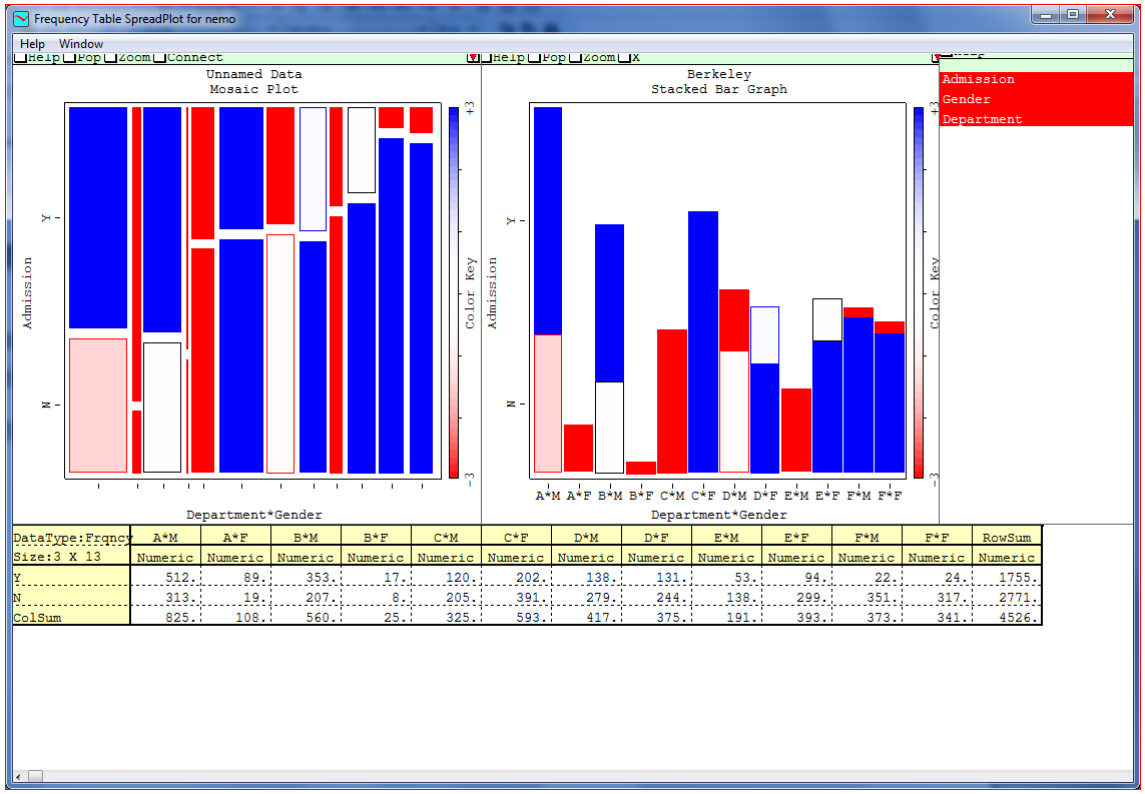
  Hay 6 departamentos divididos por género y por aceptado/rechazado

Berkeley Admissions Dataset

		Gender			
		Male		Female	
		Admission			
		Yes	No	Yes	No
Department	A	512	313	89	19
	B	353	207	17	8
	C	120	205	202	391
	D	138	279	131	244
	E	53	138	94	299
	F	22	351	24	317

- Está en data/loglinear

Spreadplot para Berkeley



DataType:Frqncy	A*M	A*F	B*M	B*F	C*M	C*F	D*M	D*F	E*M	E*F	F*M	F*F	RowSum
Size:3 X 13	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
Y	512.	89.	353.	17.	120.	202.	138.	131.	53.	94.	22.	24.	1755.
N	313.	19.	207.	8.	205.	391.	279.	244.	138.	299.	351.	317.	2771.
ColSum	825.	108.	560.	25.	325.	593.	417.	375.	191.	393.	373.	341.	4526.

- Eligiendo las variables que están en la derecha, el gráfico de mosaico cambia

  El orden en la selección de variables transmite mensajes diferentes

Ejemplo: Felicidad en función del Género/Raza

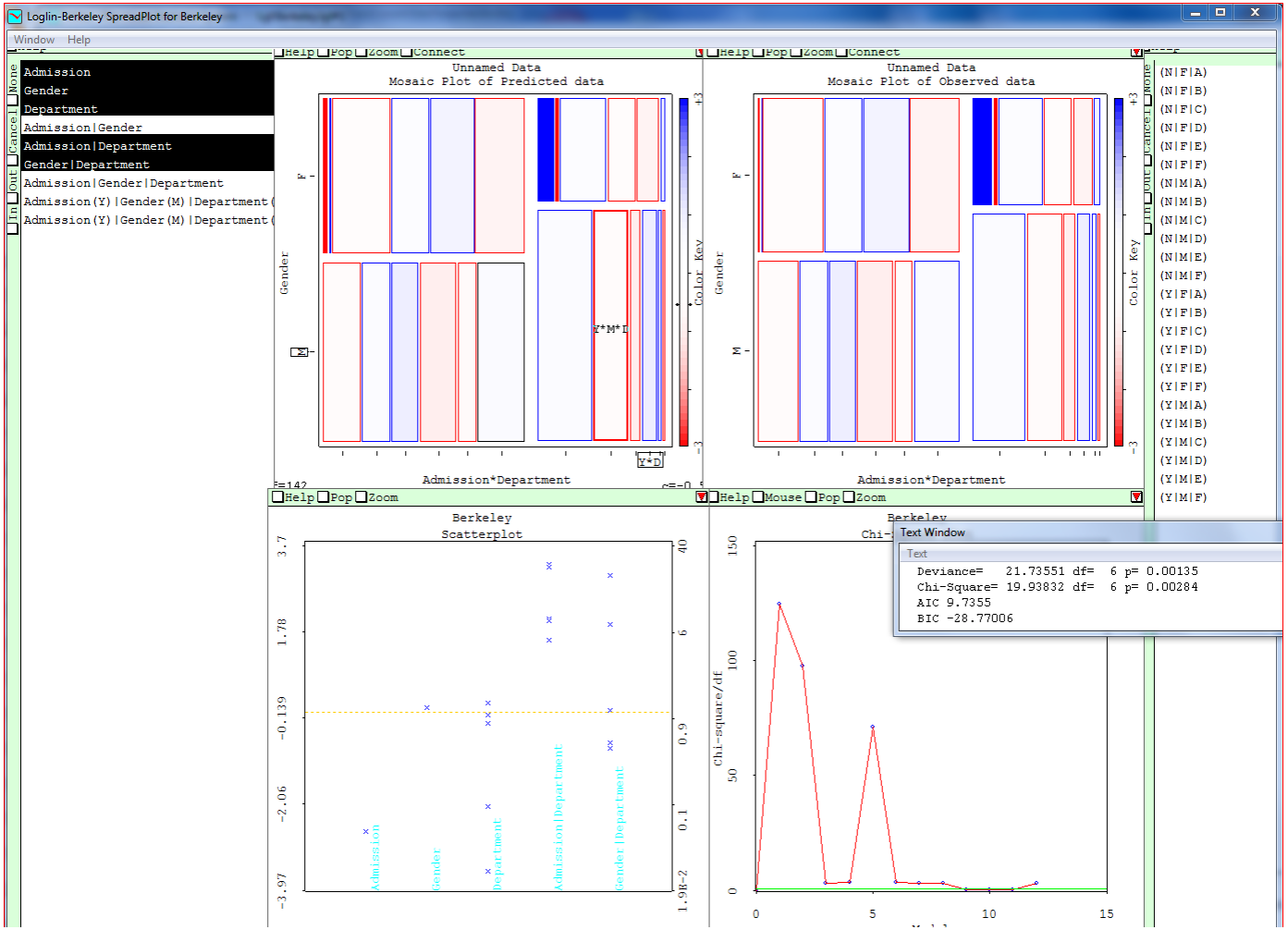
- En este ejemplo examinamos el efecto del Género y la Raza sobre la felicidad
- Este archivo se encuentra en data/loglinear/happiness.vdf

  (no happinessfreclas.vdf)

  Seleccionar sólo Happy, Gender y Race

Modelos loglineales

- Este es el spreadplot para modelos loglineales que vimos antes



A la izquierda están las combinaciones de variables para formar modelos

Funciona jerarquicamente (variables anidadas se añaden automáticamente)

Los valores de ajuste se van registrando en el gráfico de modelos

Podemos retroceder para examinar modelos antiguos

Se comparan modelos automáticamente simplemente seleccionando dos (se comprueba que estén anidados)

Hay una ventana de parámetros que indica qué elementos son responsables del ajuste

También puede funcionar no jerarquicamente

Etc., etc.

Ejemplo: Modelos loglineales para Berkeley

- Un modelo no saturado no ajusta.

  No obstante, un modelo sin la interacción de 3 orden muestra claramente que la anomalía está solo en el departamento A

- Esto lleva a un modelo no jerarquico que incluye interacción entre Género y Admisión sólo para el departamento A

  Este modelo ajusta muy bien y se interpreta del siguiente modo: No existe discriminación generalizada pero sí en el departamento A (aunque es a favor de las mujeres, no en contra)

Ejemplo: Modelos loglineales para Felicidad

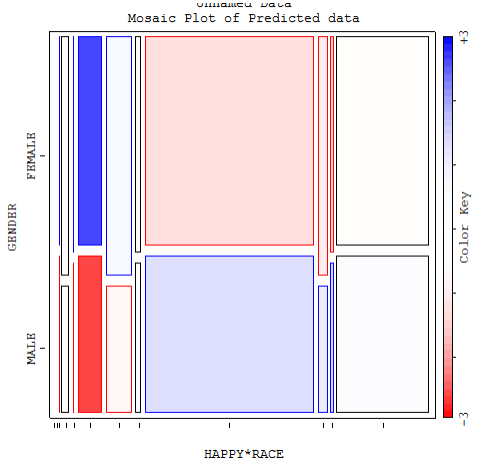
- En este modelo, resulta interesante cambiar las categorías de referencias puestas por defecto
- Si introducimos todas las interacciones de segundo orden el modelo ajusta
- No obstante, si examinamos los parámetros veremos que ninguno de segundo orden es significativo

  ¿Cómo interpretamos entonces los resultados?

  Si cambiamos las categorías de referencia es posible poner una diferente y entonces los parámetros son más fácilmente interpretables

  En concreto, podemos usar not very happy como categoría de referencia en la variable Happy. Eso muestra que la felicidad parece ser cosa de hombres!

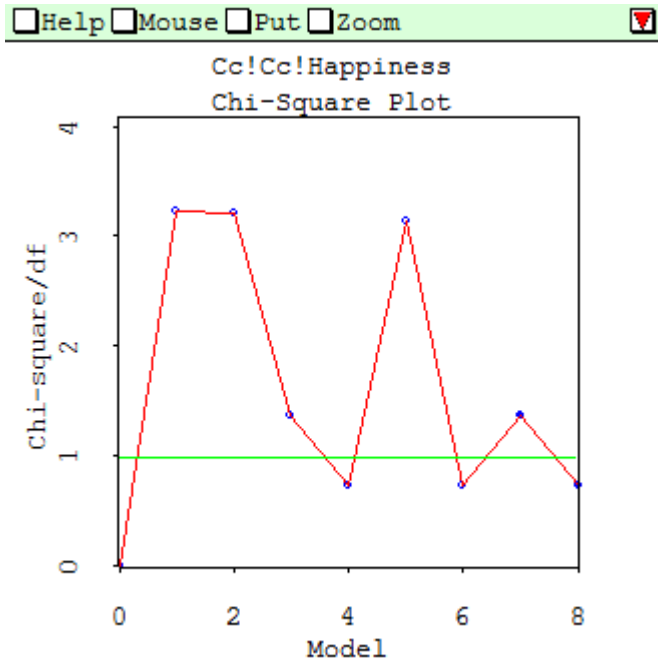
- Por otro lado, si quitamos la interacción entre Felicidad y Género llama la atención en el gráfico de mosaico la diferencia entres las mujeres y los hombres de raza blanca que declaran no ser muy felices



Un modelo con ese término ajustado individualmente ajusta muy bien

Comparación de modelos

- En el ejemplo de felicidad es posible ver que un modelo que no incluya la interacción entre Happy y Gender ajusta casi bien (Deviance 13.4 con 9 g.l; p=0.14402)
- Sin embargo, si se incluye este término el ajuste es bueno (Deviance 4.875 con 6 g.l.; p=0.55994)
- Resulta interesante comparar estos dos modelos. Esto se puede hacer seleccionando los dos modelos en la ventana de historia



Ejemplo: Modelo Logit para Sexo

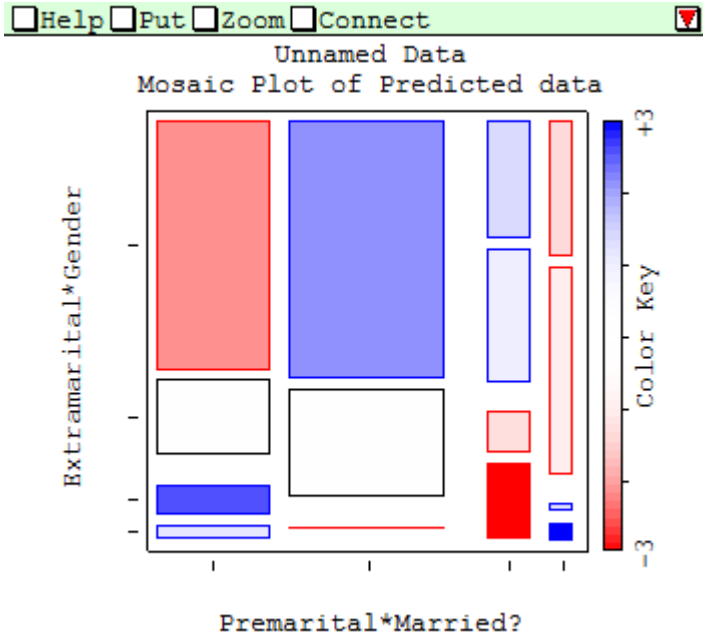
- Este ejemplo está en Freq/Sex.vdf
- Después de empezar con el modelo loglinear y tener el spreadplot a la vista se escribe en el listeners

  (send current-model :dependent variable 2)

  - Esto hace que se incluyan automáticamente todos los términos que no incluyan la variable Married?

Interpretación de parámetros

- La interpretación de parámetros en análisis loglineal es un tanto complicada
  - No hay que interpretar términos que están anidados dentro de otros de nivel superior
  - El coeficiente es una tasa que depende de las categorías de referencia
  - La ventana de parámetros ofrece información sobre el coeficiente, su significación y su cálculo
- Otra forma de interpretar es desconectar el término y examinar el gráfico de mosaico



Datos numéricos univariados

Histogramas

- Los histogramas son una de las representaciones gráficas más básicas

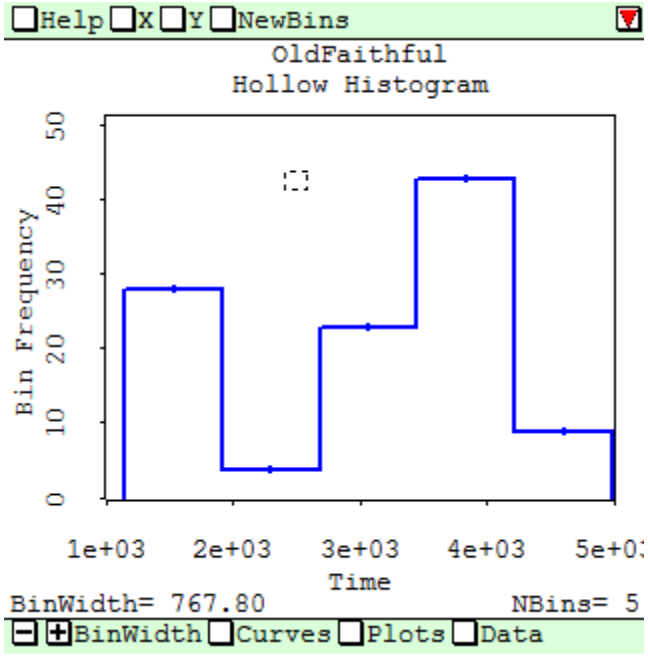
  No obstante, los histogramas son problemáticos en dos aspectos

  - El número de barras: Diferente número de barras produce diferentes histogramas
  - El punto de origen: Diferentes puntos de origen produce diferentes histogramas

Ejemplo: Old Faithful

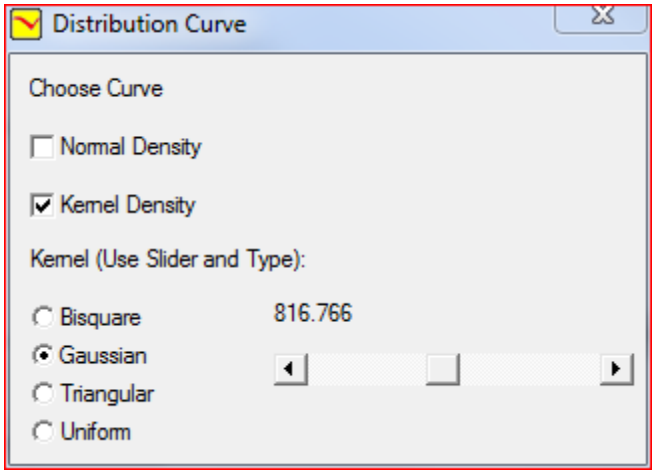
- Estos datos están en Data/general/oldfaith.vdf

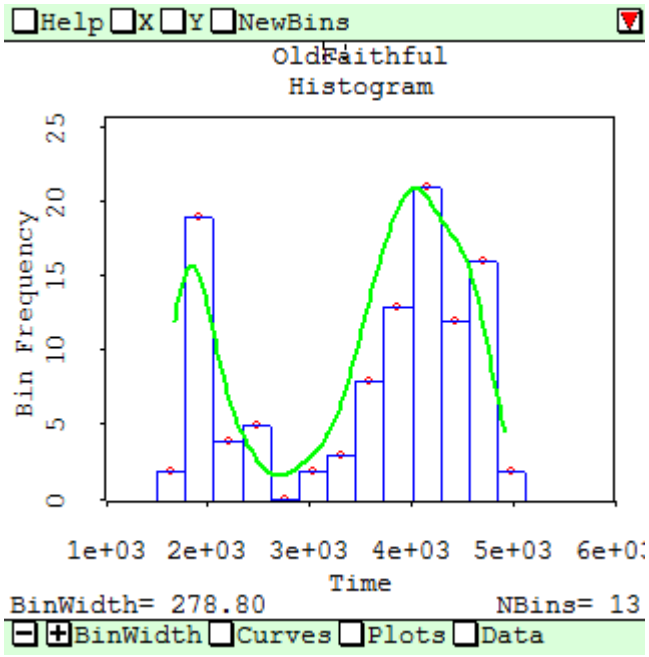
  El comando Hollow Histogram produce el siguiente histograma



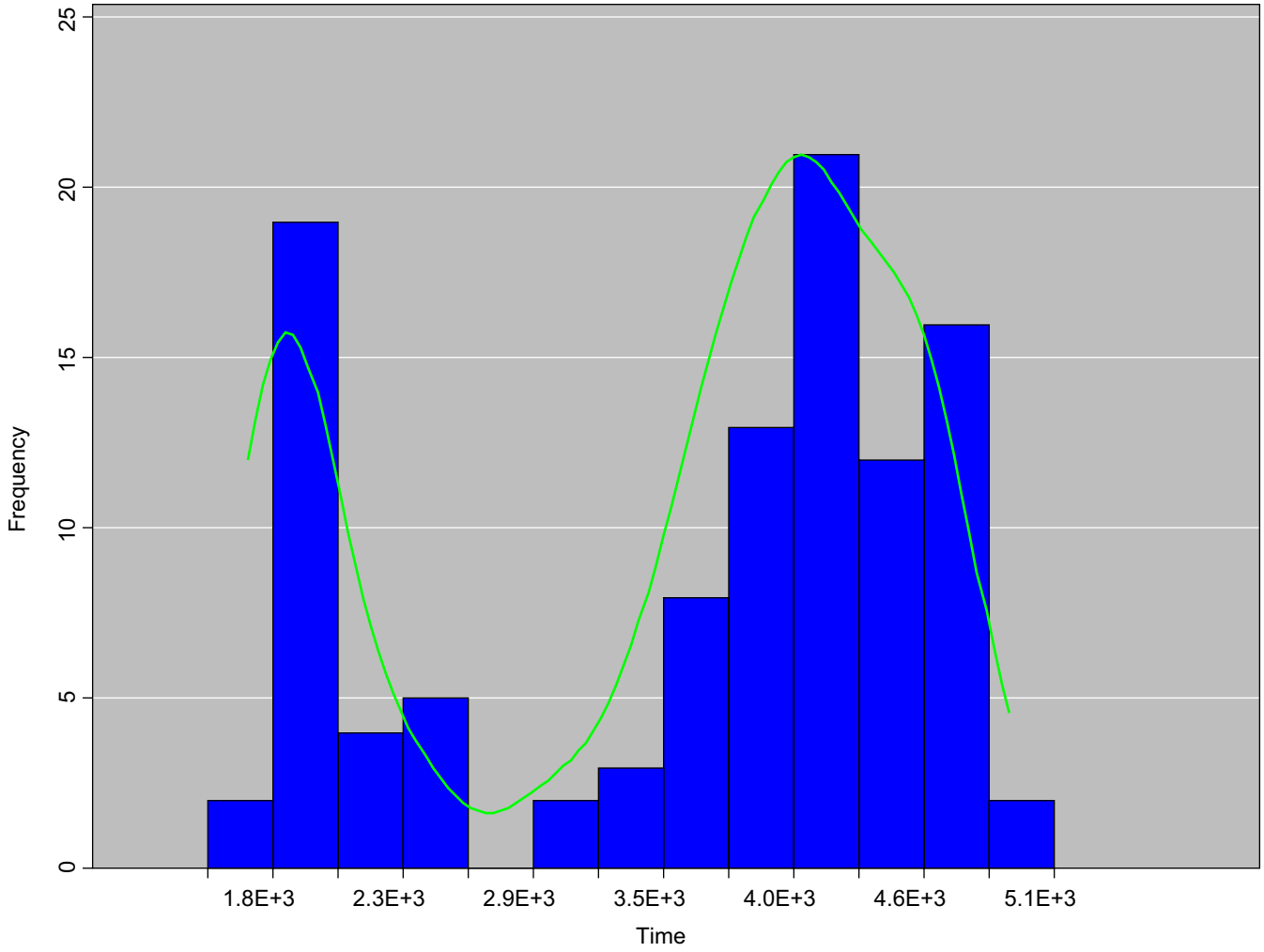
  Usando Plots y BinWidth se puede jugar para ver el efecto

- En curves podemos añadir una curva y con el slider controlar como de suave es el ajuste





- Del resultado podemos crear una figura apta para publicación



Ejemplo: Bigmac

- En data/regress/bigmac.vdf hay un archivo de datos con precios, salarios y otras variables de capitales del mundo

  Si se exploran una por una se puede encontrar que algunas de ellas son unimodales aunque asimétricas y otras son más bien bimodales e incluso trimodales

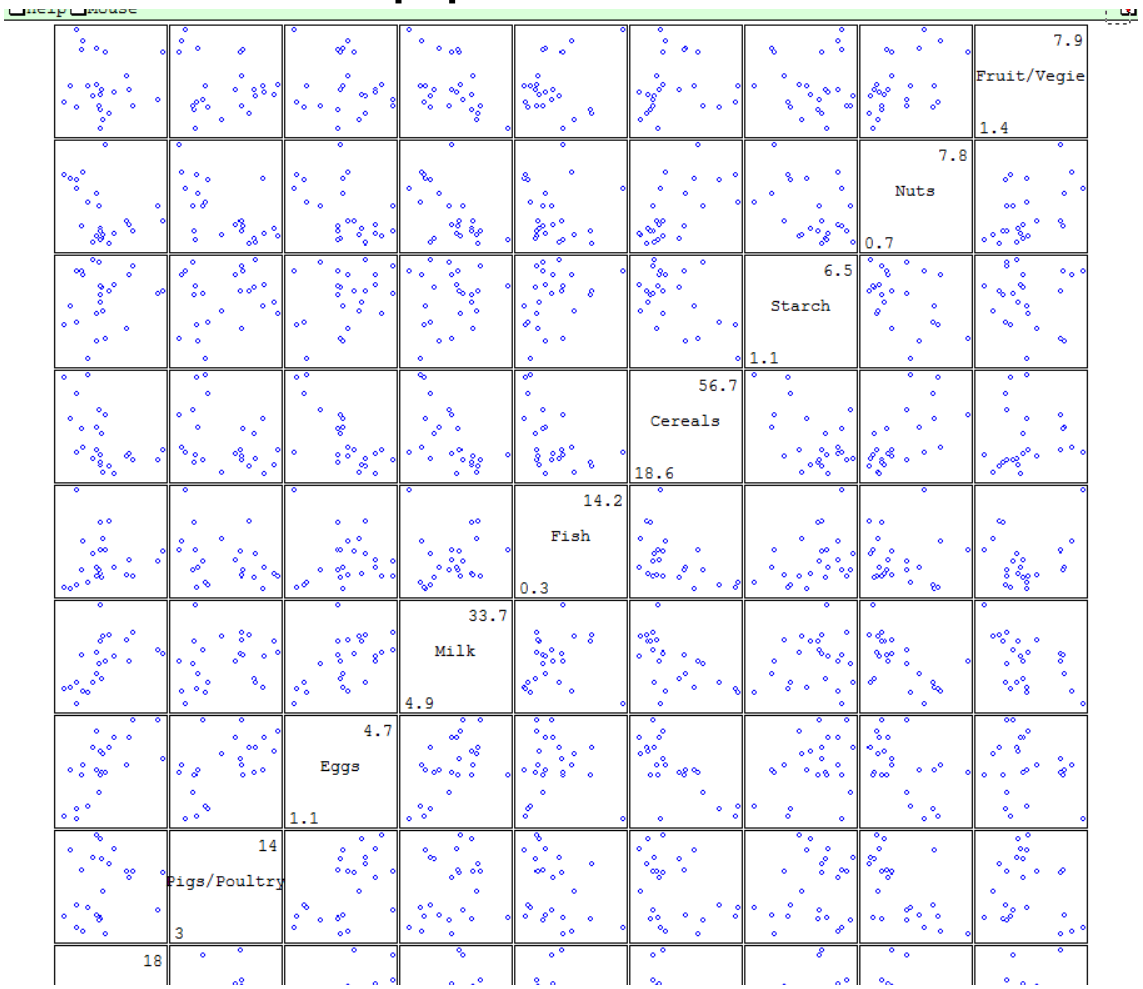
Datos numéricos bivariados

Matrices de diagramas de dispersión

- Permiten mostrar todos los diagramas de dispersión simultaneamente
- Es posible identificar valores destacados, relaciones curvilineas, etc.

Ejemplo: Proteinas en Europa 1970

- Este archivo está en data/corresp/protein.vdf



Sólo utilizaremos las variables de proteinas (Meat hasta Fruit/Vegie)

- ¿Qué países destacan?
- ¿Hay valores extremos?

- Preguntas
  - ¿Qué país consume más dieta mediterranea?
  - ¿Qué país consume peor dieta en total?

Datos numéricos trivariados

Spinplots

- Los gráficos Spinplots permiten visualizar tres dimensiones

  Al rotar suavemente, la ilusión de espacio es más fuerte que al estar parados

  Cuando se detecta una perspectiva interesante se puede parar

- Algunas rotaciones interesantes

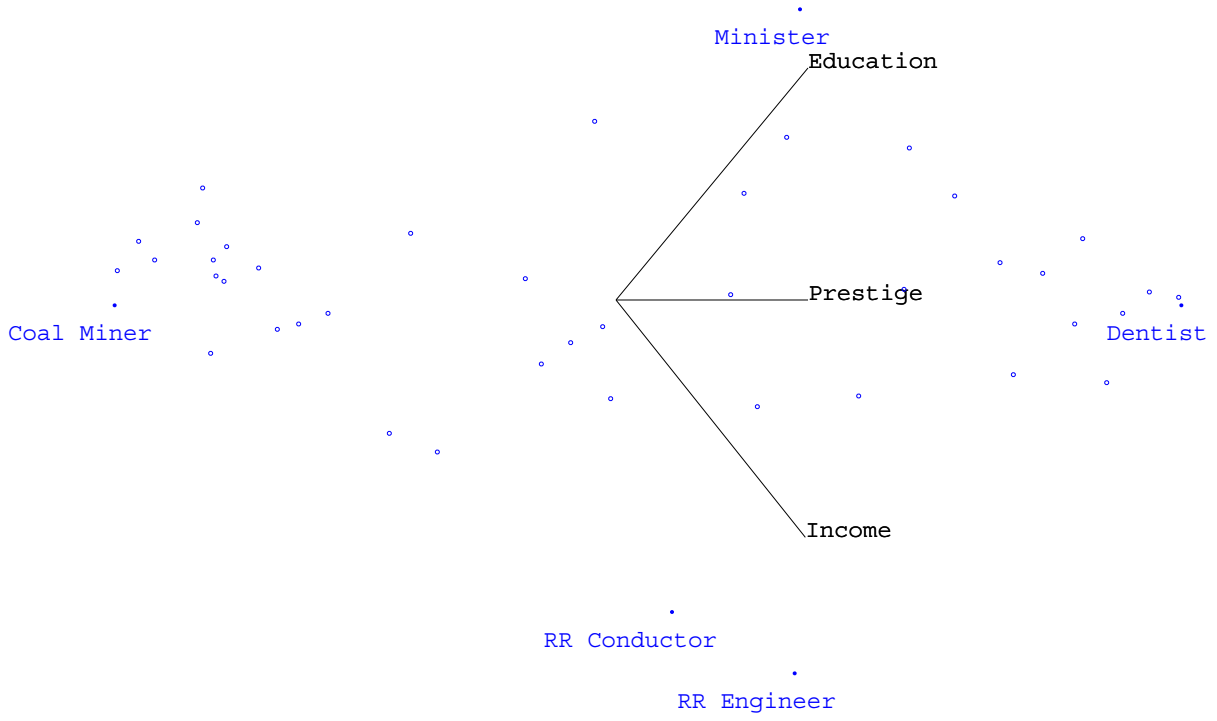
  Componentes principales: Rotar para abarcar el máximo de varianza

  Regresión: Usando este gráfico es posible estudiar dos predictores

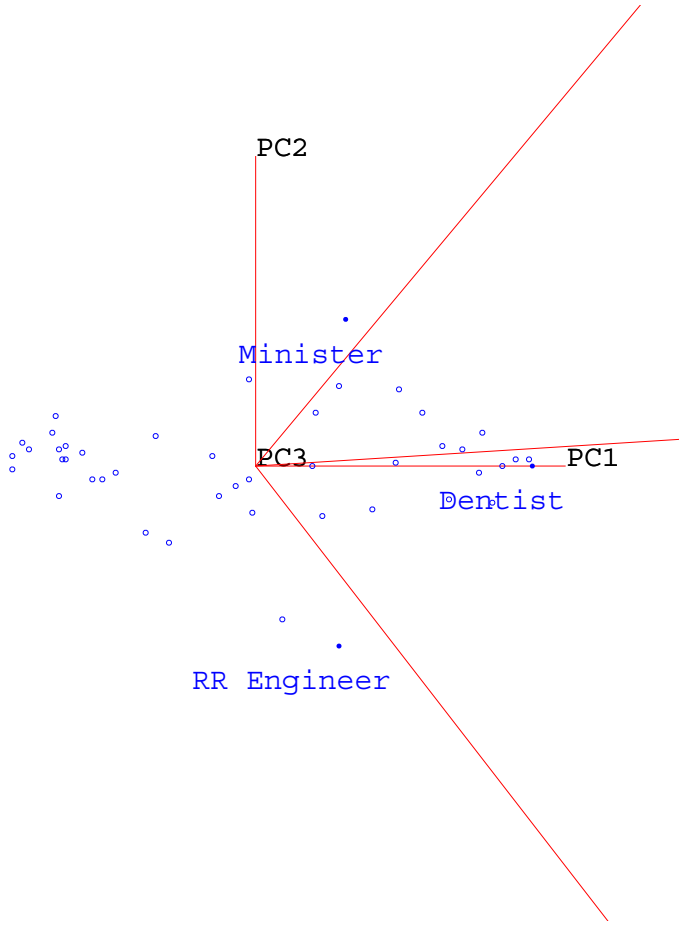
  Observar regularidades en las observaciones, valores extraños, etc.

Ejemplo: Componentes Principales en Jobs

- Rotando el Spinplot para Income, Prestige y Education obtenemos esto

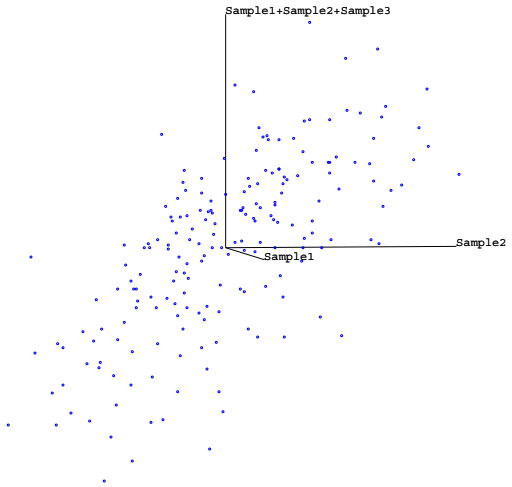


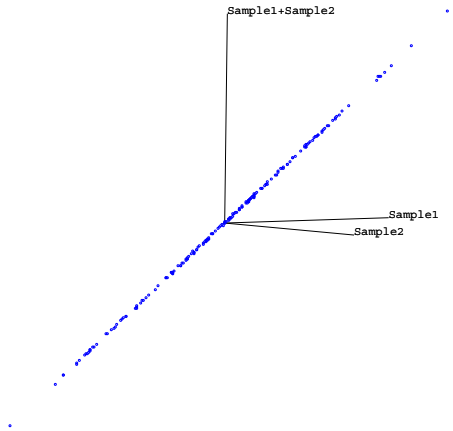
- Haciendo el análisis utilizando el programa de Principal Components de ViSta obtenemos esto



Regresión

- Usando rotaciones pueden estudiarse la relación de dos variables con una tercera.





Aquí se visualizan dos variables normales aleatorias y la suma de ambas

Observando regularidades

- En mi página web hay una animación sobre Randu, el generador de números aleatorios que no eran tan aleatorios

  Hay una demostración en ViSta que puede obtenerse usando el comando File>Load Edit y buscando el archivo data/general/Randu.vaf

  Eso produce una demostración de que en una dimensión o en dos dimensiones el generador no parece problemático pero en tres sí

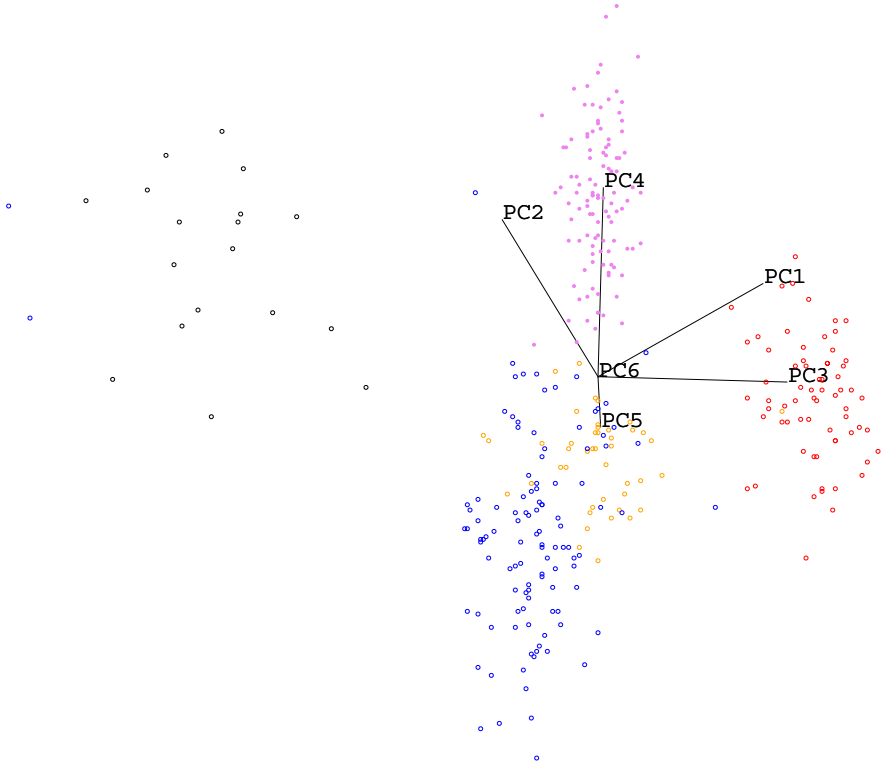
Datos numéricos multivariados

Técnicas

- Grand tours
- Componentes Principales, Biplots y Análisis Factorial
- Cluster jerarquico

Tours

- Un tour es un gráfico en movimiento diseñado para estudiar la distribución conjunta de datos multivariados



- Los tours son creados haciendo una secuencia de proyecciones en dos dimensiones de datos multidimensionales

  Esto puede servir para encontrar relaciones que implican muchas variables

  La idea es buscar proyecciones que sean interesantes a partir de la visualización de muchas de ellas en una especie de película o animación

  Cook and Swayne (2007) y el software que lo acompaña es probablemente la mejor referencia

- Los métodos de búsqueda que hay son:
  - Grand Tours: Los espacios visualizados son elegidos aleatoriamente
  - Projection Pursuit Tour: Las búsquedas están guiadas por un algoritmo que sugiere projecciones interesantes
  - Busqueda manual: El usuario elige la proyección y permite explorar un espacio cercano a una proyección que parece interesante

- Muchas de las proyecciones obtenidas en los Tours están muy relacionadas con técnicas tradicionales que son vistas de modo numérico:
  - Los biplots hechos a partir de componentes principales
  - Análisis discriminante está conectada con la proyección que mejor separa las medias de los grupos
  - El analisis de correlación canónica y la regresión múltiple multivariada también producen proyecciones interesantes

Ejemplo: Crimes

- El Tour plot en ViSta utiliza como índice a maximizar uno derivado de los componentes principales
- Utilizando el archivo de datos de crime que está en data/princomp y el orbiting plot podemos ver un tour

  Un ejemplo de las cosas que podemos fijarnos es en puntos que se desplazan de una manera diferente a los demás

  Hawai es uno de esos puntos

  Para interpretar ese punto podemos hacer un diagrama de cajas paralelo y seleccionar Hawai

  Podemos ver que Hawai es especial porque los niveles de Asalto son bajos pero los de otros crímenes no

Componentes Principales y Biplots

- Los componentes principales son proyecciónes que muestran donde los datos están más extendidos (mayor varianza)

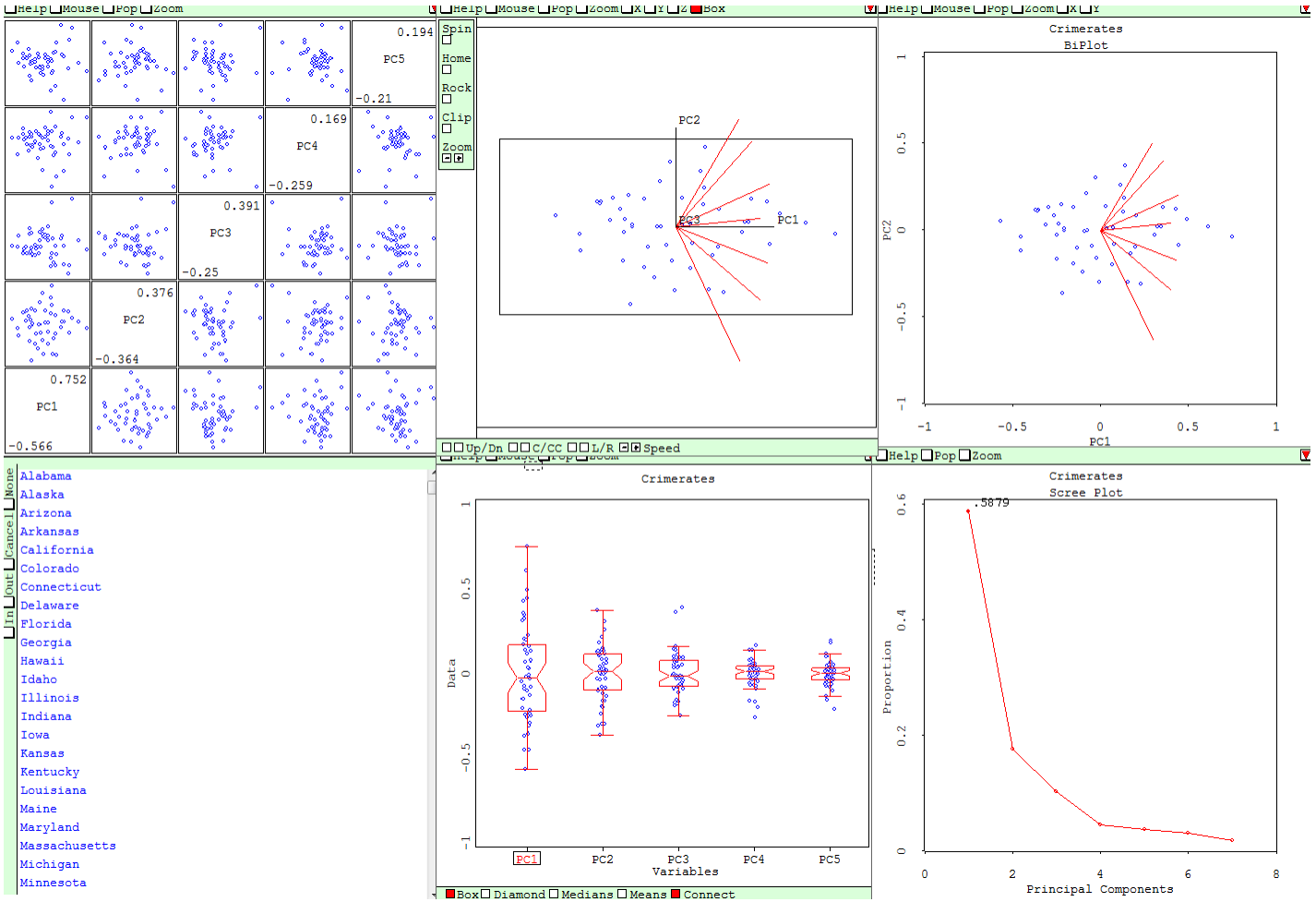
  Las proyecciones sucesivas son ortogonales entre sí y buscan donde los residuales respecto de las dimensiones previas están más extendidos

- En ViSta, podemos visualizar los resultados de un análisis de componentes principales en un spreadplot que incluye un biplot

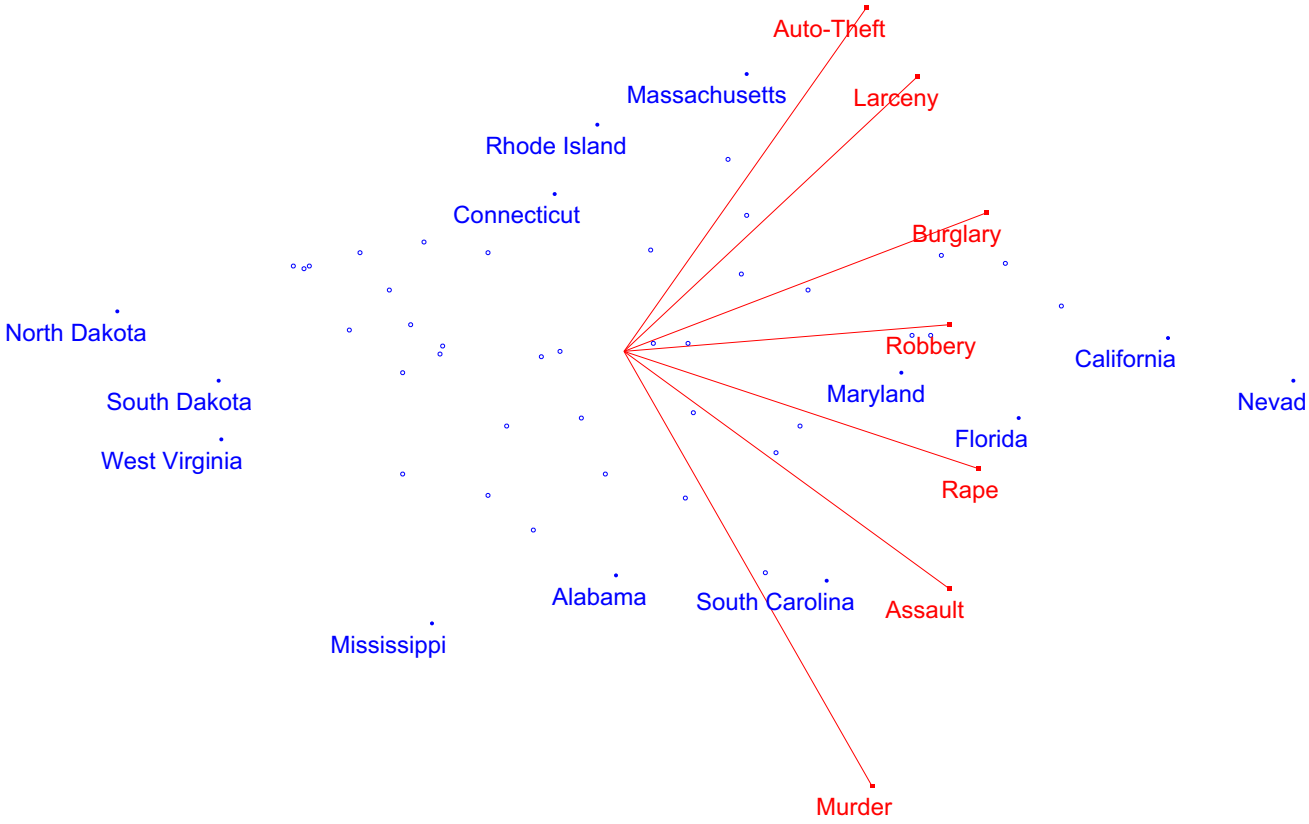
  El análisis puede basarse en *correlaciones* (cuando la varianza de las variables está en diferentes escalas) o *covarianzas* (la varianza de las variables está en la misma escala)

Ejemplo: Crímenes en Estados en USA

- Spreadplot



- Biplot

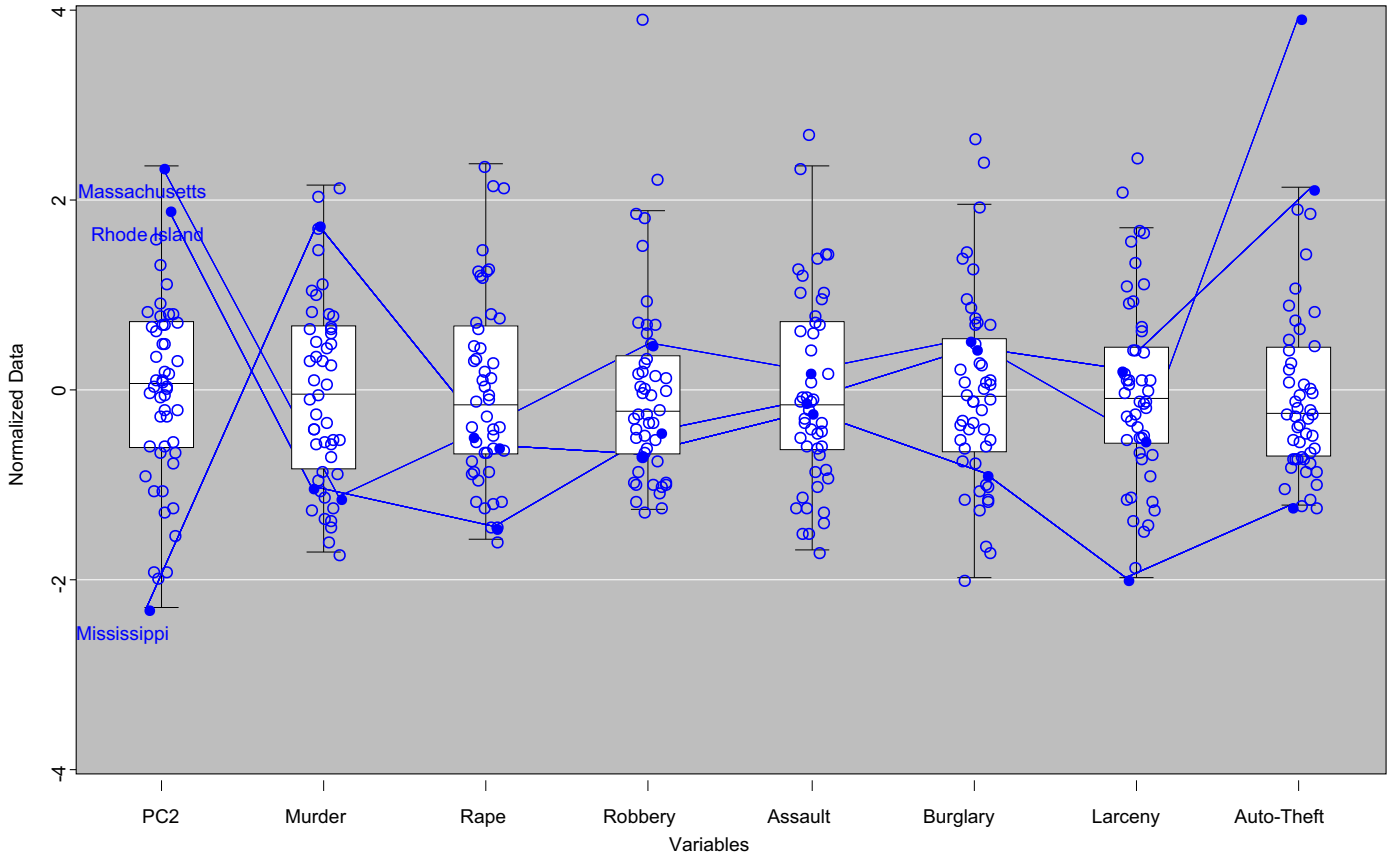


Este es un biplot del primer y el segundo componente principal

Podemos valorar relaciones entre variables y posiciones individuales

- En un Biplot vemos
  - Eigenvalores, proporción de varianza explicada por cada componente y acumulada
  - Eigenvectores: Proyecciones de los vectores de las variables sobre los componentes principales
  - Puntuaciones en los componentes: Es la matriz de puntuaciones en los componetes por la raíz cuadrada de los eigenvalores. La proyección sobre el PC nos da idea de como una observación está explicada por ese PC

- Interpretación



El PC1 está relacionado con volumen, pero el PC2 distingue perfiles de estados (crímenes contra las personas v. crímenes contra la propiedad).

En esta trasparencia se pueden ver un par de casos característicos

Ejemplo: Proteínas

- Este ejemplo está en data/corresp/protein.vdf.
- Un análisis de componentes principales nos muestra una gran cantidad de detalles de interés

  Alimentos que se suelen ir acompañados

  Países que destacan en esos alimentos

  Valores extremos, etc.

Cluster jerárquico

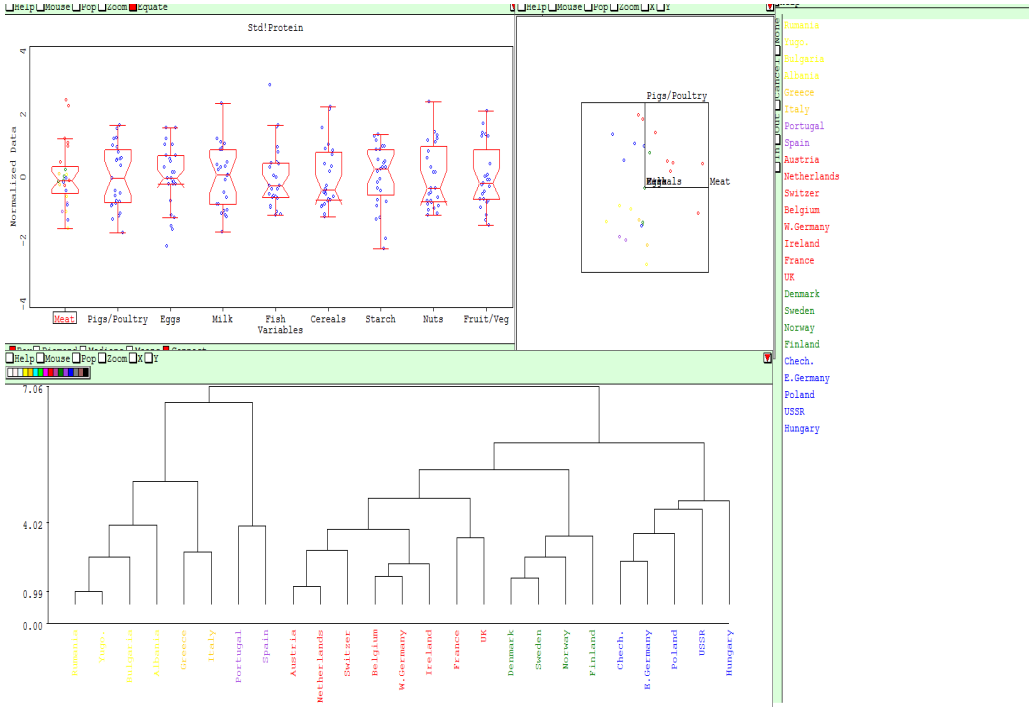
- Podemos analizar los datos de proteínas utilizando dendogramas calculados utilizando el módulo de Cluster jerárquico

  Antes de empezar el análisis, pasaremos los variables a puntuaciones z

  En la figura siguiente hemos usado distancias euclidianas y linkage completo

  Al estar conectado el dendograma con los otros gráficos es posible explorar el significado de los grupos más fácilmente

- Spreadplot para cluster jerárquico



Este ejemplo usa linkage completo y es necesario estandarizar los valores antes

El resultado muestra fundamentalmente 4 grupos de países que pueden ser subdivididos posteriormente

Los grupos identificados por colores se pueden grabar como un archivo de datos para hacer análisis posteriores (no cerrar el spreadplot, sólo minimizar)

Ejemplo: Horas de trabajo, Precios y Sueldos en ciudades

- Los datos están data/cluster/Cities.vdf
- Los datos tienen un par de valores perdidos que se puede utilizar mediante Impute Missing Data en el menú de Data
- A partir del análisis cluster con linkage completo se puede dividir los datos en 6 grupos de ciudades

  Explorando los grupos se pueden ver las características que tienen y las diferencias

  También se pueden detectar algunas anomalías

Datos perdidos

El desafío de los datos perdidos multivariados

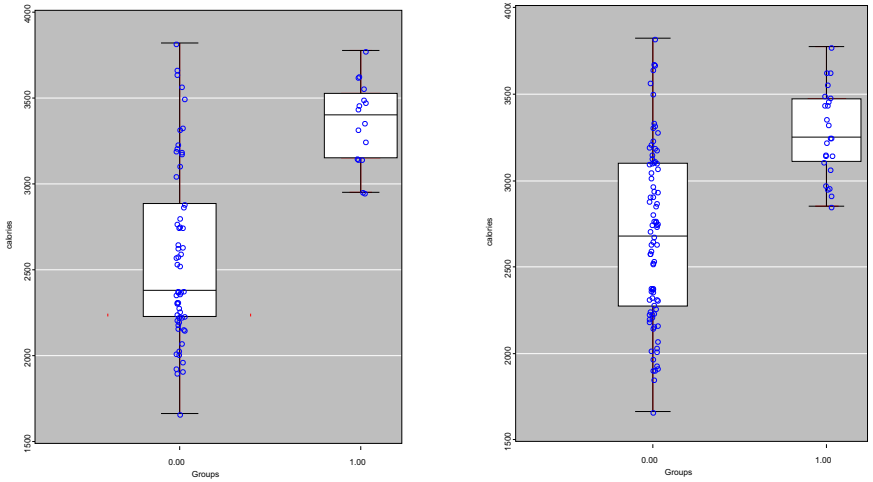
- Cuando hay datos perdidos, los gráficos se encuentran con muchos problemas
  - Diferentes gráficos univariados pueden tener diferentes casos
  - A medida que intentamos representar más variables, el número de casos completos puede disminuir mucho
- Con gráficos ligados, los gráficos pueden ser muy confusos
  - Datos que se iluminan en unos gráficos pero no en otros
  - Conexiones que no son posibles
- MANET tiene gráficos adaptados a esta situación
- En ViSta hay dos aproximaciones a visualizar datos con valores perdidos:
  - Patrones de datos perdidos y valores
  - Visualización después de hacer imputaciones de datos

Visualización de Patrones de Perdidos

- En principio, una visualización interesante es tratar los valores perdidos/ observados como una variable de grupo y comparar los valores en las otras

  No obstante, la variable partida también puede tener valores perdidos por lo que siempre existe la duda acerca de si nos estamos perdiendo algo

  En los dos gráficos anteriores, se ha clasificado en perdido/observado la variable Litfemale y se visualiza Calorías. En la derecha se han imputado valores (encuentra las diferencias)



Ejemplo: Mundo95

- Usaremos como ejemplo el archivo de World95 que está en data/missing

  ViSta tiene un spreadplot que está especializado en examinar datos que tienen valores faltantes

- Ese spreadplot está enfocado a ver los *patrones de datos faltantes*

  Los datos faltantes en una variable a menudo no vienen solos

  Hay varias variables que coinciden en tener los mismos valores faltantes

  Conocer su asociación puede ser de gran interés para entender los problemas que hay en ellos

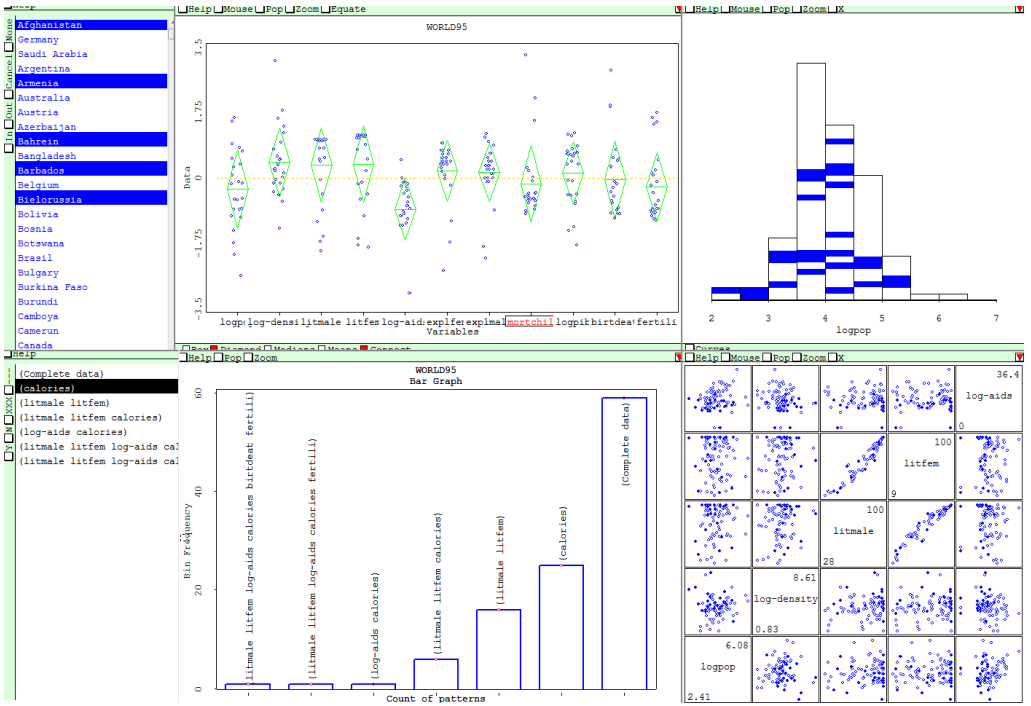
- El spreadplot para valores faltantes tiene este aspecto

  Hay que entender que cada gráfico intenta mostrar el máximo de información disponible en cada caso

  El gráfico de puntos paralelos muestra todos los valores observados para el patrón de datos seleccionado

  El histograma muestra todos los datos observados de la variable mostrada, y cuando se selecciona el patrón se ilumina mostrando los valores observados

  La matriz de diagramas de dispersión muestra todos los valores observados en las variables mostradas en él en ese momento



- En este ejemplo, es fácil ver que el patrón Litmale Litfemale está asociado con valores concretos de las variables

Imputando los datos

- Imputar datos con valores razonables tiene muchas ventajas desde el punto de vista de la visualización

  Los gráficos vuelven a ser completos

  Se puede tener una idea de qué valores estaban perdidos

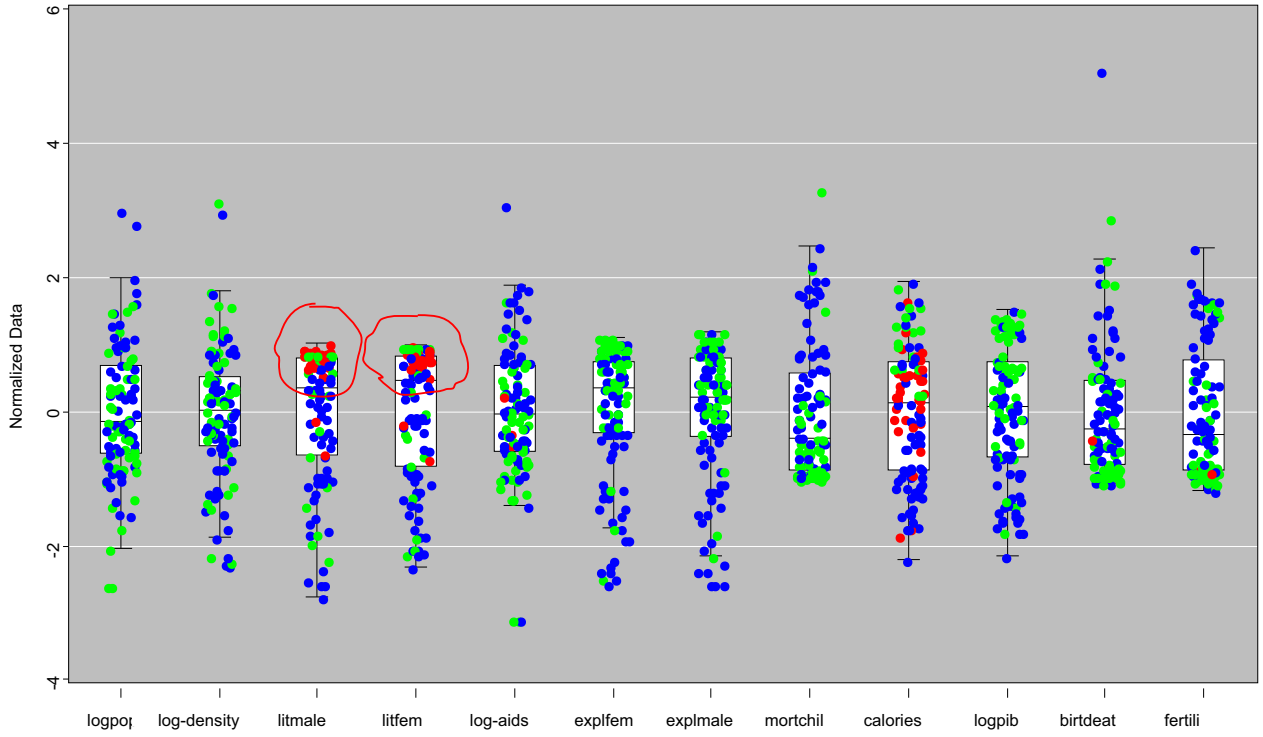
- También tiene sus inconvenientes

  Técnicamente, imputar valores razonables puede ser costoso

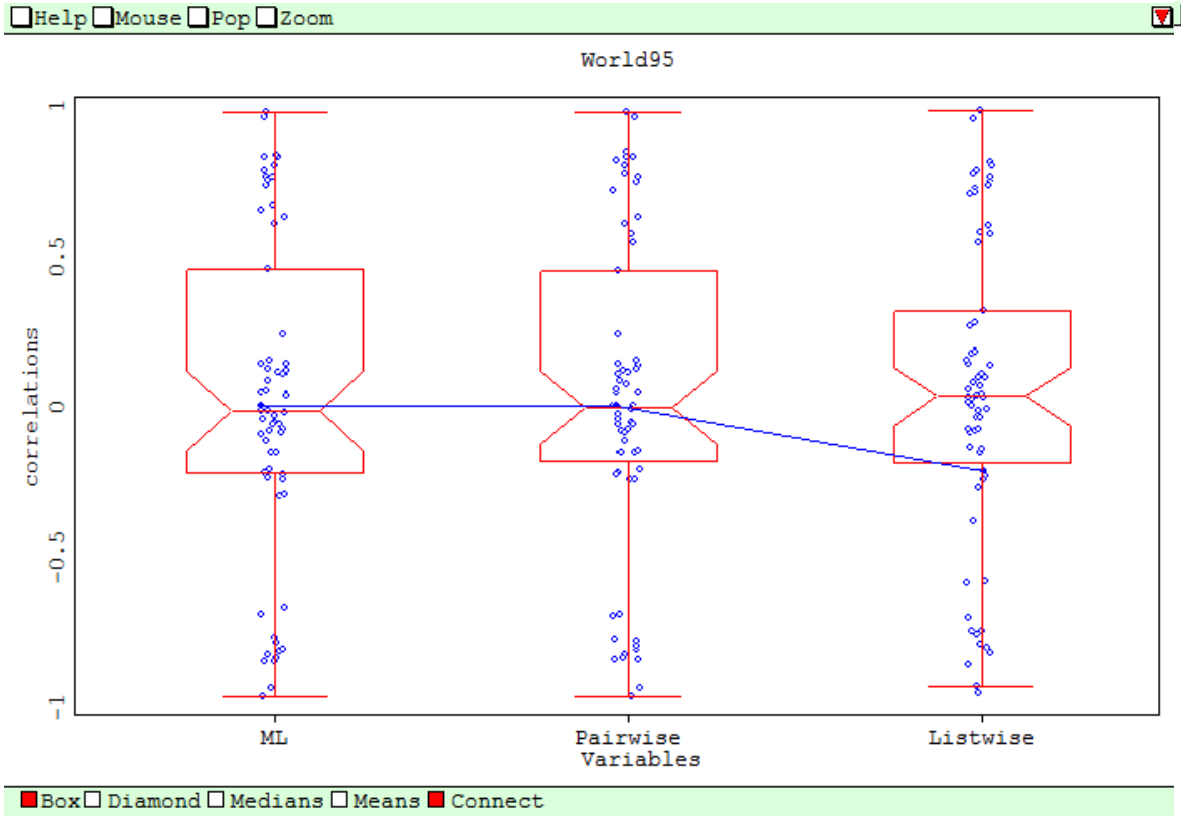
  - Seleccionar el método
  - Cumplir los supuestos (si se utiliza un método basado en predicciones lineales)
  - El coste computacional puede ser excesivo

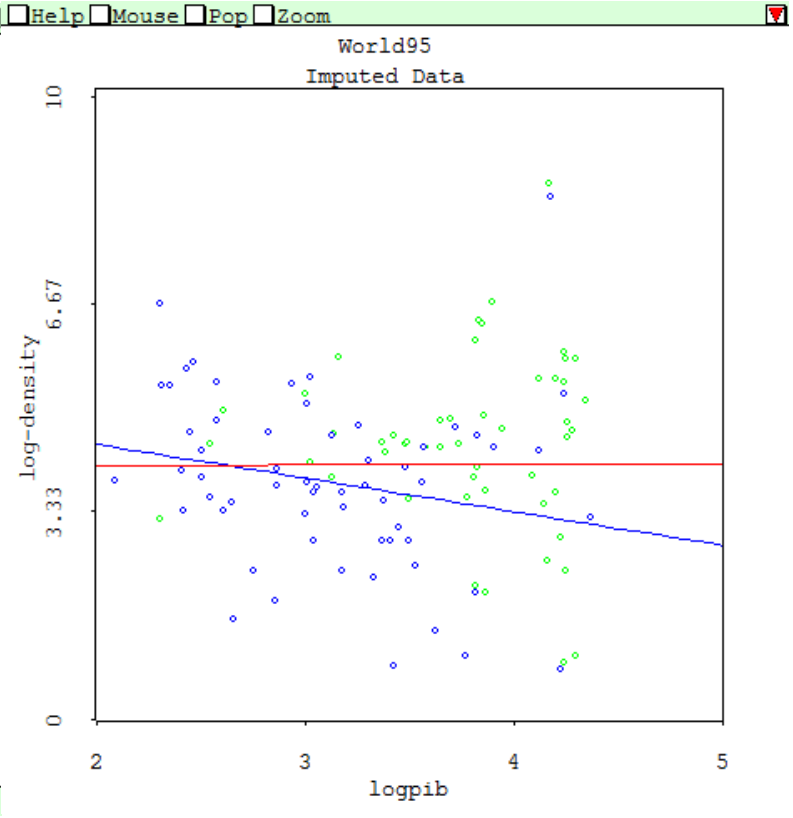
Ejemplo: World95

- El comando Impute Missing Data en el menú Data usa el algoritmo EM basado en mínimos cuadrados para estimar medias y correlaciones entre las variables. A partir de ese resultado se puede hacer imputación simple y visualizar



- Otro aspecto interesante es examinar si hay variaciones en las correlaciones



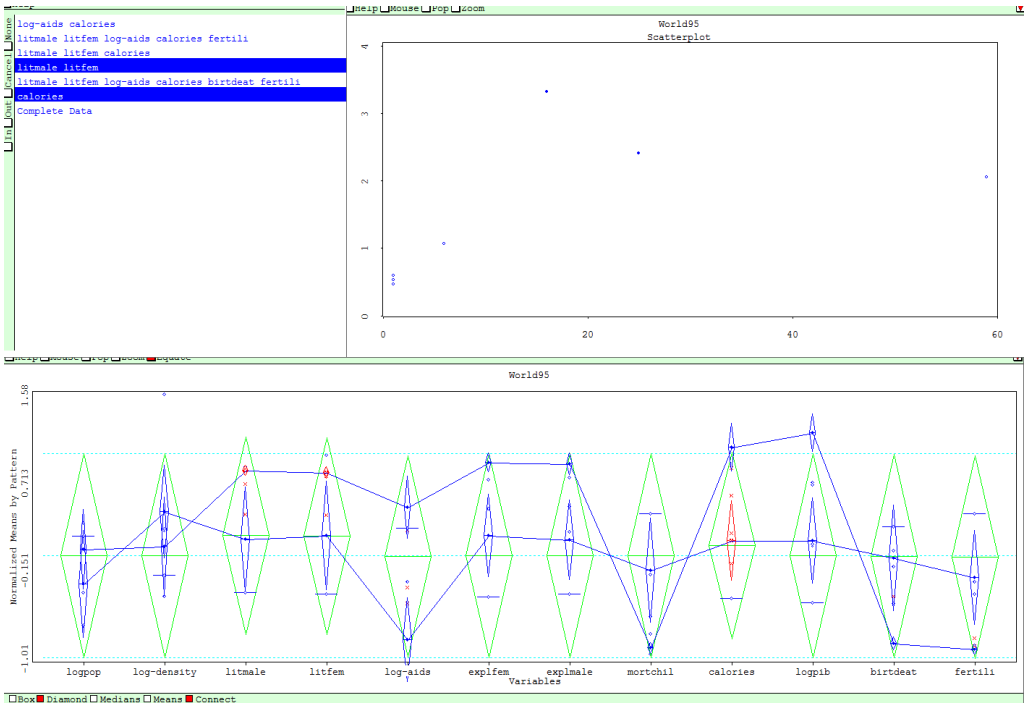


Grandes diferencias entre las correlaciones nos alertan de lugares en los que los datos perdidos han causado mayores estragos

Un ejemplo muy llamativo es el de data/missing/marijuana.vdf

- El boxplot de puntos se puede simplificar para mostrar solamente la información por patrones
  - El gráfico de arriba derecha está basado en el test de MCAR de Little

    Ese test compara las diferencias entre las medias observadas y las medias estimadas por max. verosimilitud por patrón y las suma



El gráfico de arriba muestra las diferencias por patrón de datos perdidos y sirven de indicador de como los valores perdidos en un patrón están asociados con los valores observados en otras variables, y, al imputar, las medias estimadas son diferentes de las observadas

Ejemplo: Titanic

- Este ejemplo está en data/missing/titanic2.vdf

  En estos datos se muestran datos acerca de la supervivencia del titanic

  La variable Edad falta en muchos de los datos

  Una visualización muestra que Edad está asociado con no viajar en primera o segunda clase

    Imputar los datos y hacer el gráfico de patrones (send current-model :visualize-patterns) muestra este resultado bastante claramente

Escalamiento multidimensional

Recuperando posiciones a partir de distancias

- ViSta tiene dos módulos sobre escalamiento multidimensional:
  - MDS promediado métrico es apropiado para distancias métricas e ignora cuando hay varias matrices de distancias (las promedia)
  - Multidimensional scaling: Admite distancias ordinales entre los objetos así como puede utilizar diferentes matrices siempre y cuando sean simétricas (hay programas que sí lo aceptan)

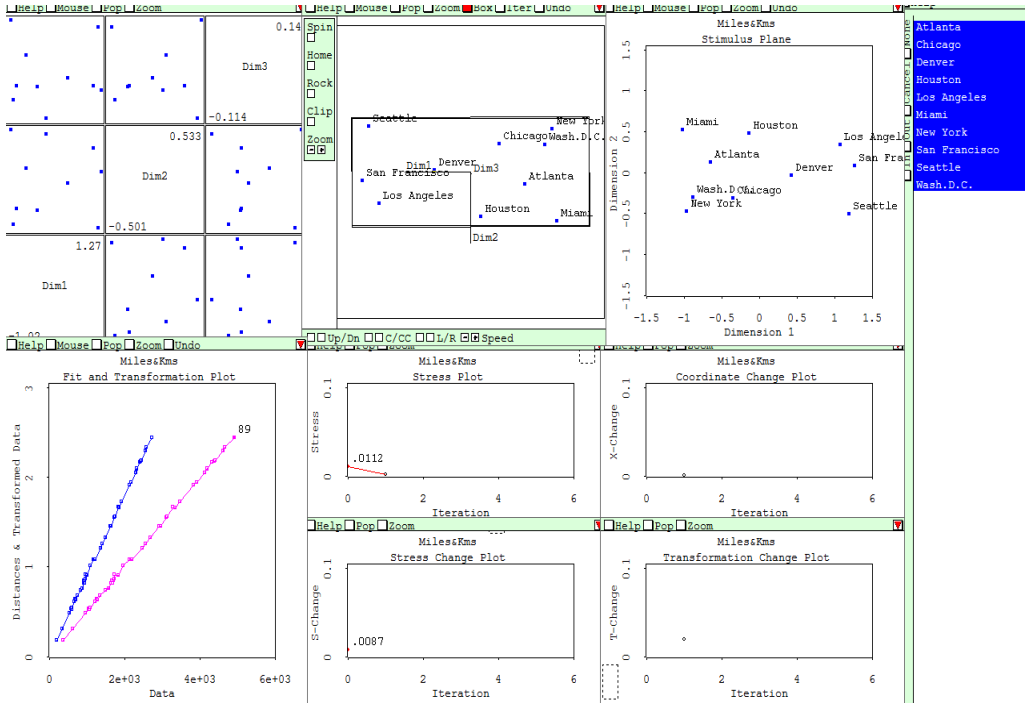
    Este módulo además ajusta distancias y no distancias al cuadrado (como hacían algunos programas más antiguos)

Ejemplo: Distancias entre ciudades

- Se trata de las distancias en kilómetros y en millas entre ciudades en USA

  Son dos matrices simétricas y equivalentes (transformación lineal)

  - El gráfico de transformaciones muestra que la transformación de las distancias orginales es lineal y semejante
  - El mapa no obstante aparece girado, usando el Spin plot se puede poner de la manera correcta



Ejemplo: Explorando la posición de los colores

- Se trata de juicios acerca de la similaridad de unos colores
- Utilizando el comando de **Metric Averaged MDS...** se aplica este método
  - A continuación se puede aplicar un número de iteraciones
  - Se pueden mover puntos para comprobar su efecto sobre el stress

Apéndices

Importando datos

- ViSta importa datos de texto

  Variables separadas por tabuladores

  Casos separados por retornos de carro

  La primera columna puede ser de etiquetas

  Valores perdidos se identifican con nil

  ViSta puede importar también datos agrupados y datos de similaridades

  Usar puntos para decimales. No usar separadores de miles

- En la carpeta Data/Import hay una serie de ejemplos que pueden ser imitados en caso de que haya problemas

  Leer también el documento Import_wisdom.txt

Guardar gráficos en formato vectorial

- ViSta no es un buen programa para crear gráficos para presentación pero a veces crear los gráficos en otros programas sería costoso así que es interesante hacerlo en ViSta
  - Muchos gráficos tienen un comando de **Save Plot as...** en el menú de la derecha (marcado con un triángulo)
  - Ese menú abre un cuadro de diálogo. Se pone un nombre y el archivo se guarda en formato .pdf y dibujado vectorialmente

    El resultado es una interpretación del gráfico, no es una versión literal en todos los aspectos
- Si se desea, se pueden manipular estos gráficos utilizando un programa de dibujo vectorial:
  - Inkscape parece que funciona bien pero Draw de OpenOffice hace un desastre
  - El programa PDF Reader (no Acrobat) pemite guardar como WMF para Word

- Ejemplos de gráficos

  Apartado de figuras de la página sobre el libro