Bivariate Descriptive Statistics

K. Gibert

Department of Statistics and Operation Research

Knowledge Engineering and Machine Learning group

Universitat Politècnica de Catalunya, Barcelona

karina.gibert@upc.edu

www.eio.upc.edu/homepages/karina

Guió

0. Two numerical variables
    1. Graphical descriptive tools
    2. Numerical descriptive tools
1. One numerical variable and one categorical
    1. Graphical descriptive tools
    2. Numerical descriptive tools
2. Two categorical variables
    1. Graphical descriptive tools
    2. Numerical descriptive tools
3. Two categorical and one numerical
4. Two numerical and one categorical

Bivariate Descriptive analysis

Compact and Informative view of the variables
RELATIONSHIP

DATA= FIT+ ERROR

Characterizacion

General Pattern **Deviations**

Structural Component Random Component

Tools

1. Graphical

   Visualitze variable's relationship

2. Numerical

   Quantify what is observed in he graphs

Cases

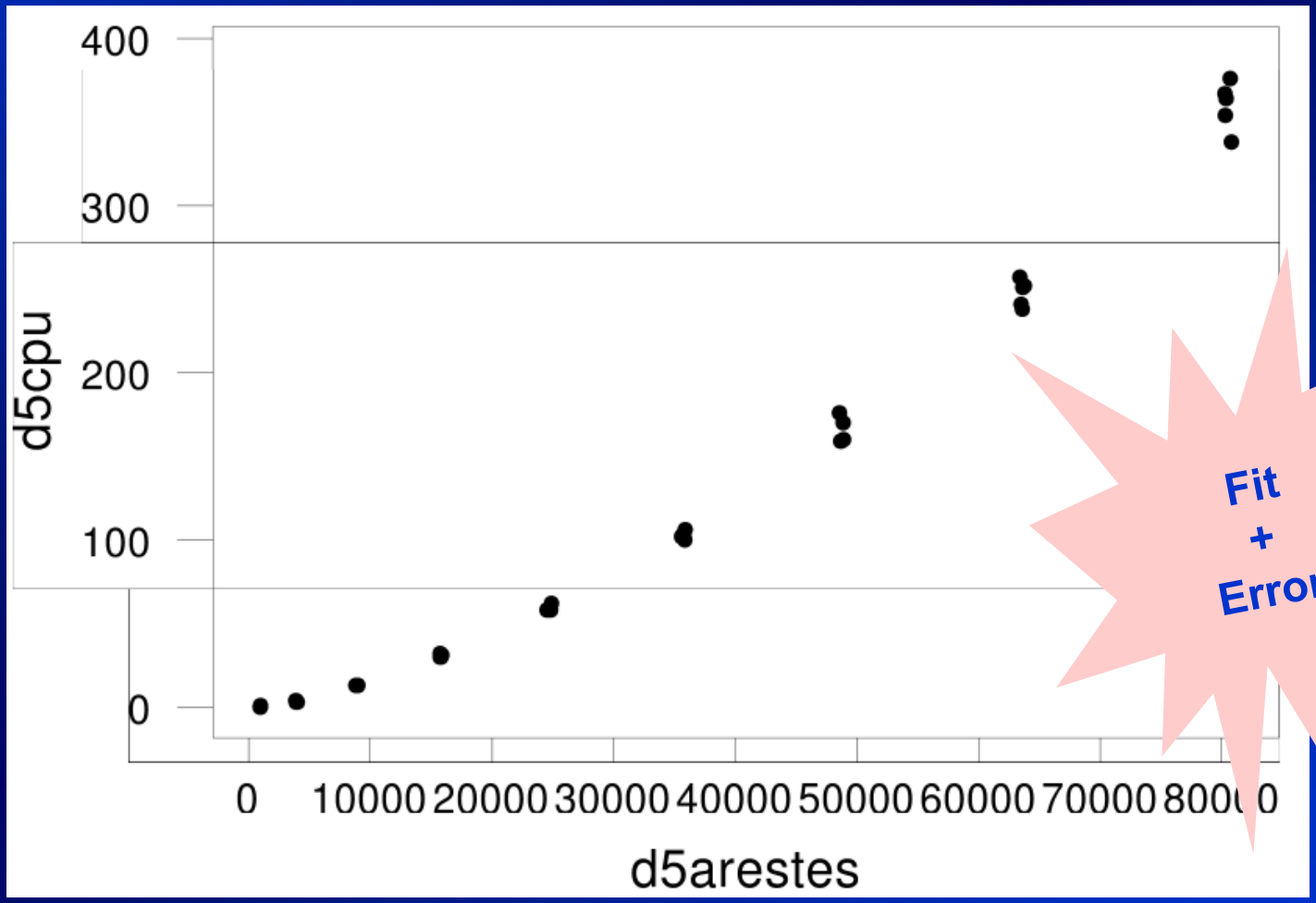
1. Two numerical variables
2. Two categorical variables
3. One categorica and one numerical

Roles of variables

- Symmetrical
- Response vs Explanatory variable

Two numerical variables

Plot



Reading a plot

1. Direction: Direct (positive) or inverse (negative)
2. Form: Central trend (structural component) *(pass a thread)*
    1. *Linear*
    2. *Polynomic*
    3. *Exponential:* ***The 70 rule****: Given a raising factor R, const… …wth takes time 70/R from Y to 2Y*

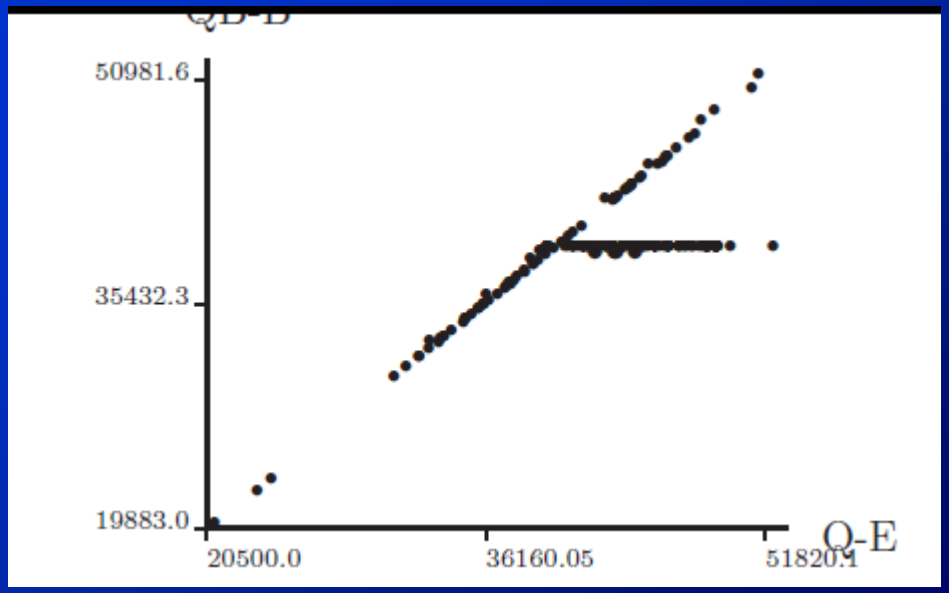
Family of equations

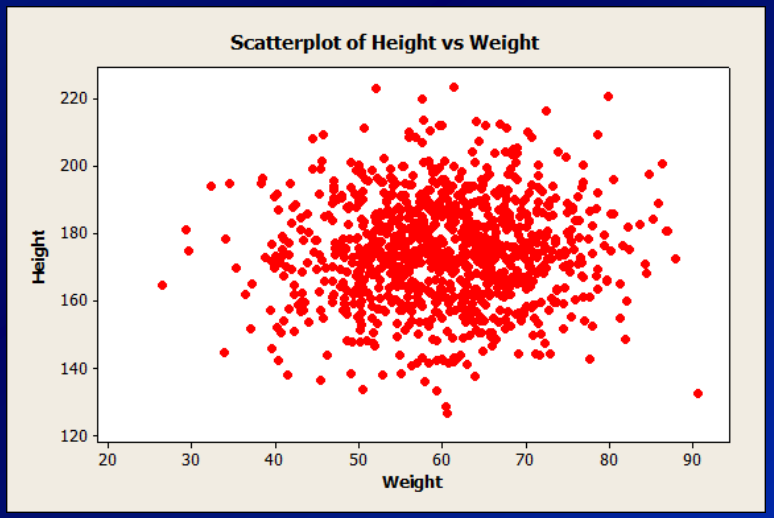
3. Intensity: Deviations arround central trend (Variability)

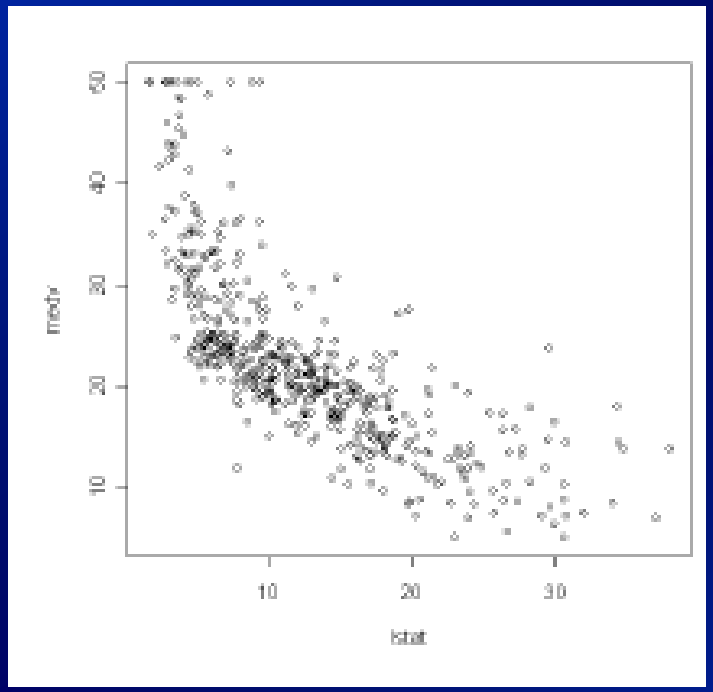
    (spaghetti vs big sausage)

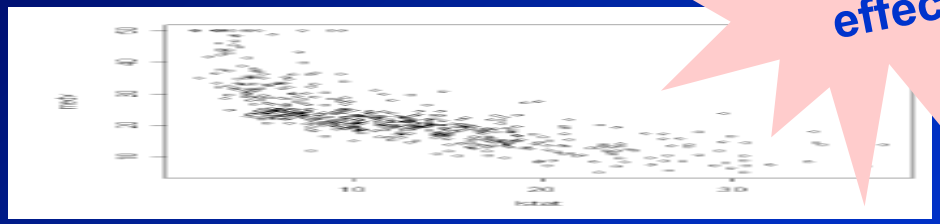
4. Trend changes
5. Ranges for X and Y
6. Bivariate outliers
7. Symmetry

Plot Case Studies



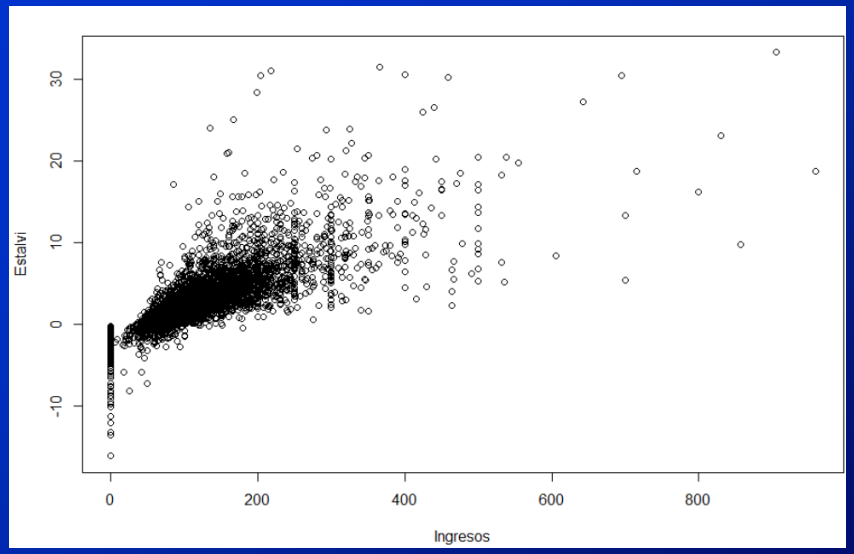


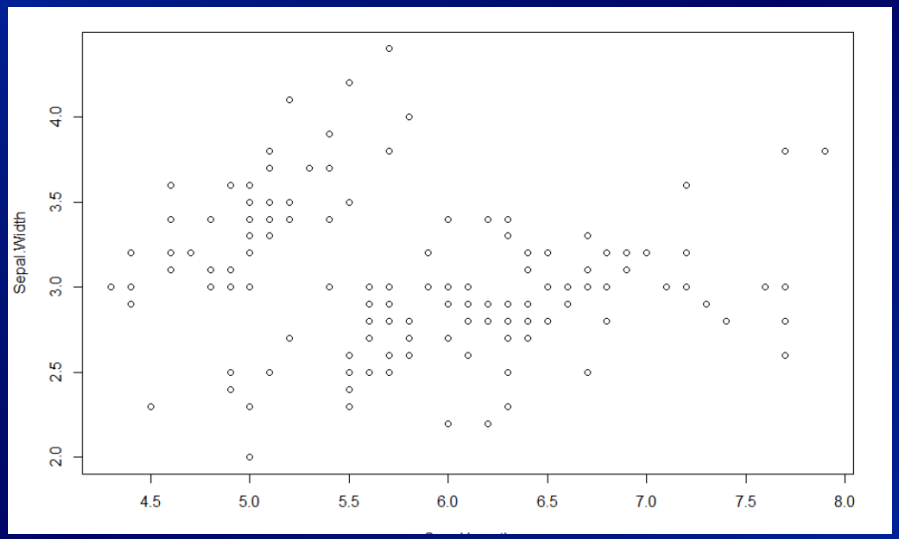


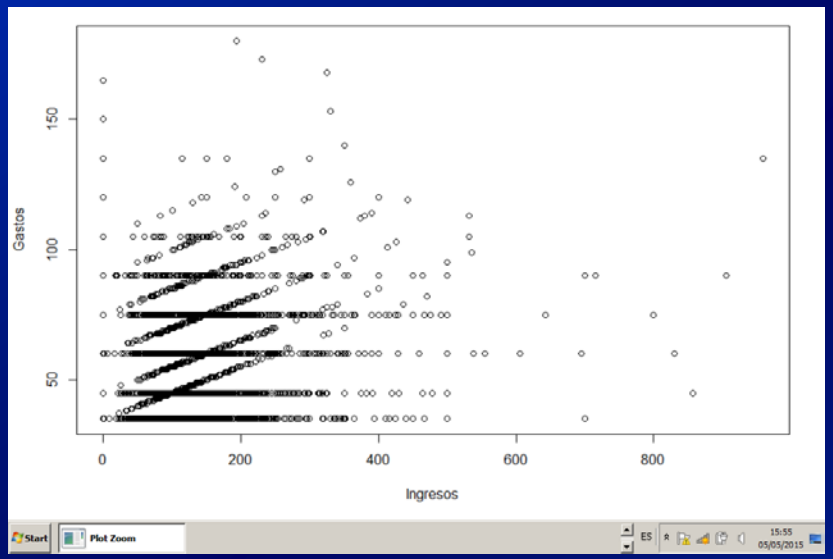


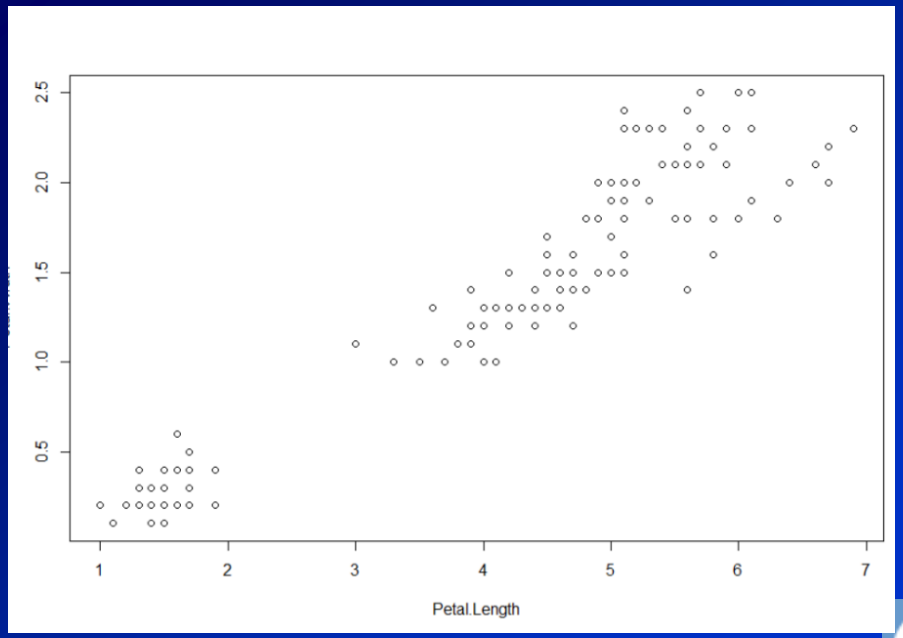
Care: Optical effects

Plot Case Studies

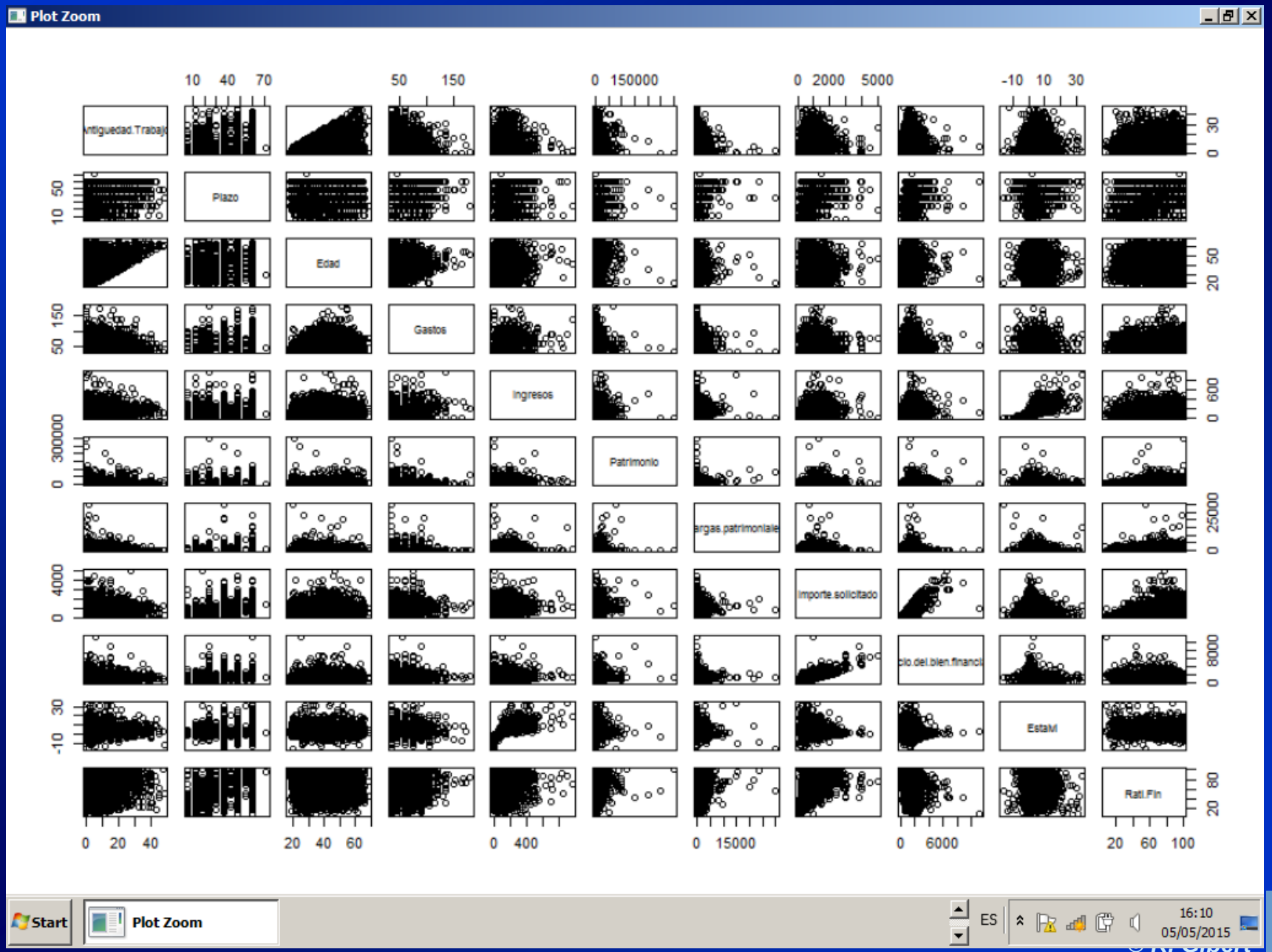








Matrix Plot


Plot Zoom
Antiguedad.Trabajo
Plazo
Edad
Gastos
Ingresos
Patrimonio
ergas.patrimoniale
Importe solicitado
cio.del.bien.financi
EstaMI
Ratl.Fin
Start
Plot Zoom
ES
16:10
05/05/2015

Reading a plot

1. Direction Direct (positive) or inverse (negative)
2. Form: Central trend (structural component) *(pass a thread)*
   1. *Linear*
   2. *Polynomic*
   3. *Exponential:* ***The 70 rule****: Given a raising factor R, const… …wth takes time 70/R from Y to 2Y*
3. Intensity: Deviations arround central trend (Vari…bility)

   (spaghetti vs big sausage)
4. Trend changes
5. Ranges for X and Y
6. Bivariate outliers
7. Symmetry

Family of equations

Association between numerical variables

Quantify by Correlations test

Only linear relationships

Consider general coefficients if required

Sheffer Generalized coefficient

Interpreting Correlation

- Covariance: Dimensional

  depends on data measurement units

- Correlation: Adimensional

  Assumes Linear relationship

  Sign: Indicates Direction of relationship (>0: positive)

  Magnitude: Indicates Intensity

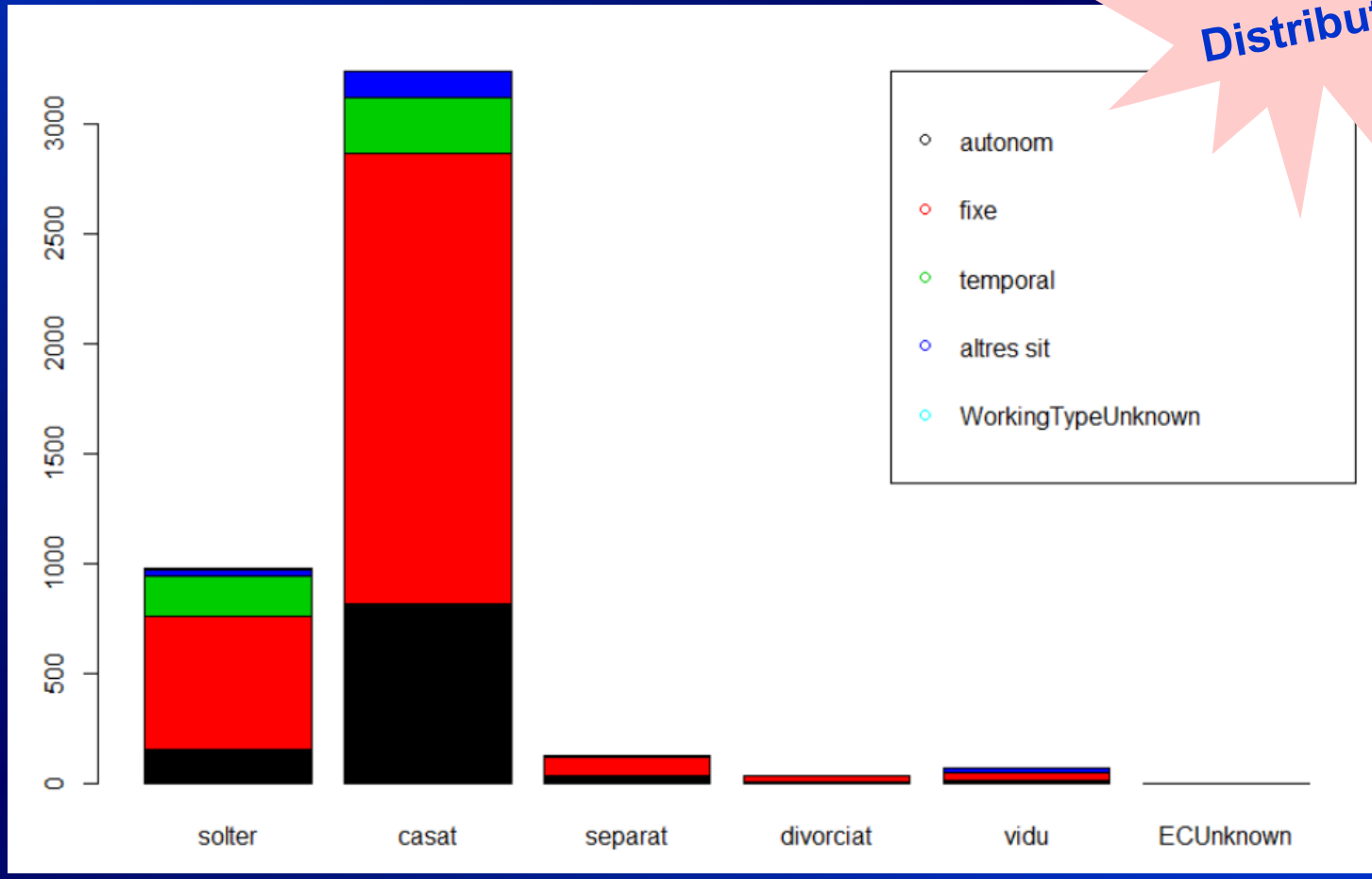
  |1| perfect linear association

  0: Linear independence

Two categorical variables

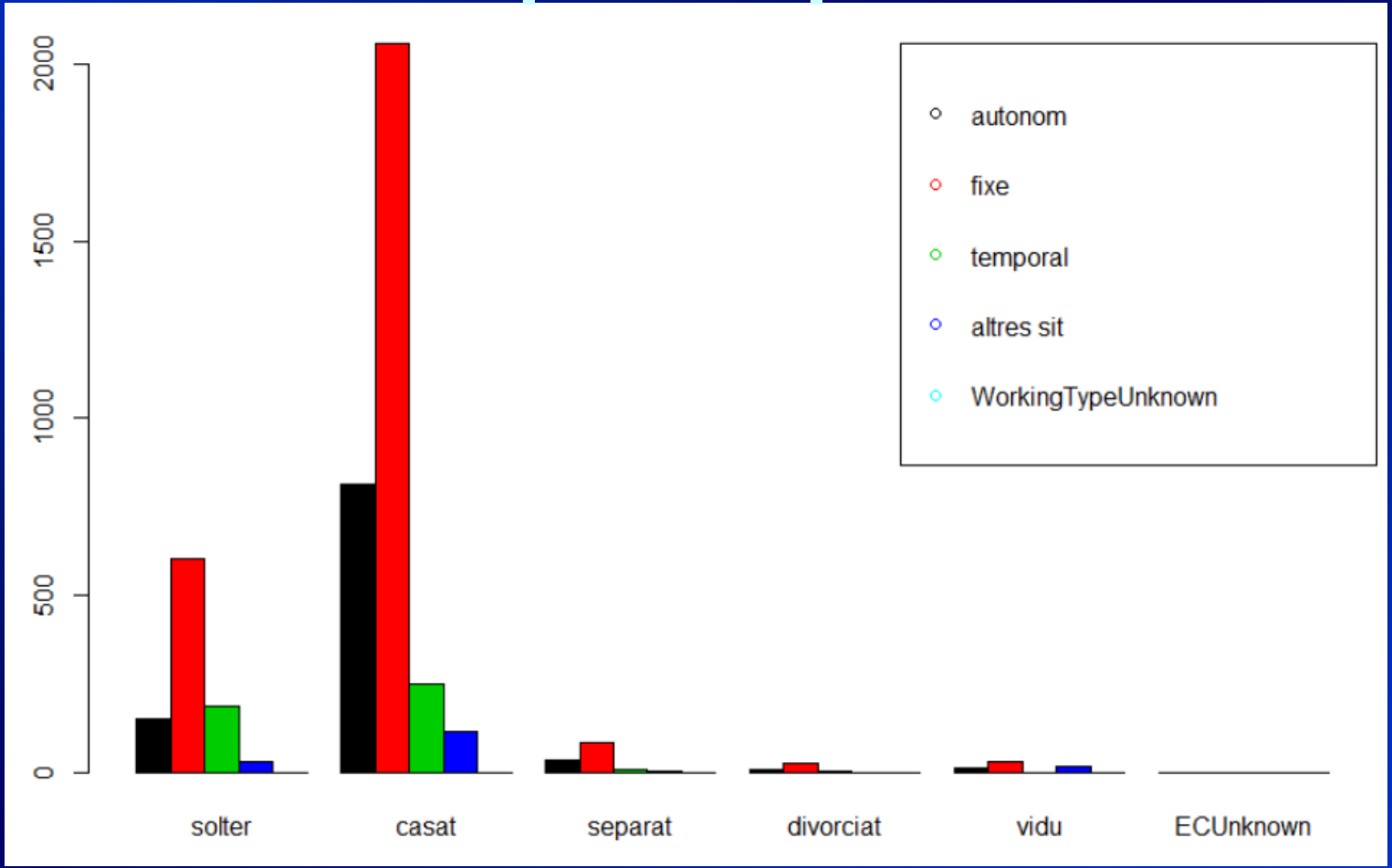
Multiple barplot

Joint Distribution



Two categorical variables

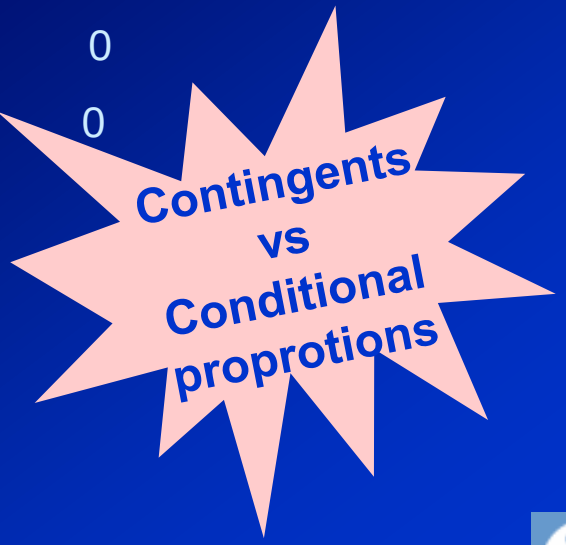
Multiple barplot



Contingency tables

(Cross Tables)

	Estado.civil					
Tipo.trabajo	solter	casat	separat	divorciat	vidu	ECUnknown
autonom	154	815	34	7	13	1
fixe	605	2056	84	28	33	0
temporal	188	252	8	3	1	0
altres sit	29	118	4	0	20	0
WTUnknown	2	0	0	0	0	0



Contingency tables

(Margins)

	Estat_civil						
Vivenda	solter	casat	vidu	separat	divorciat	Total	*Row %*
lloguer	174	723	11	50	15	973	*21.9%*
escriptura	167	1839	50	38	12	2106	*47.4%*
contr_privat	26	212	3	4	1	246	*5.5%*
ignora_cont	1	18	0	0	1	20	*0.4%*
pares	507	238	0	30	7	782	*17.6%*
altres viv	98	208	3	8	2	319	*7.2%*
Total	973	3238	67	130	38	4446	
Columns %	*21.9%*	*72.8%*	*1.5%*	*2.9%*	*0.9%*		

Assessing association between categorical variables

The chi2 independence Test

Missperformance if nkj<5

Care with Simpson's Paradox

Assessing association between categorical variables

The Simpson's Paradox

Apparently independent

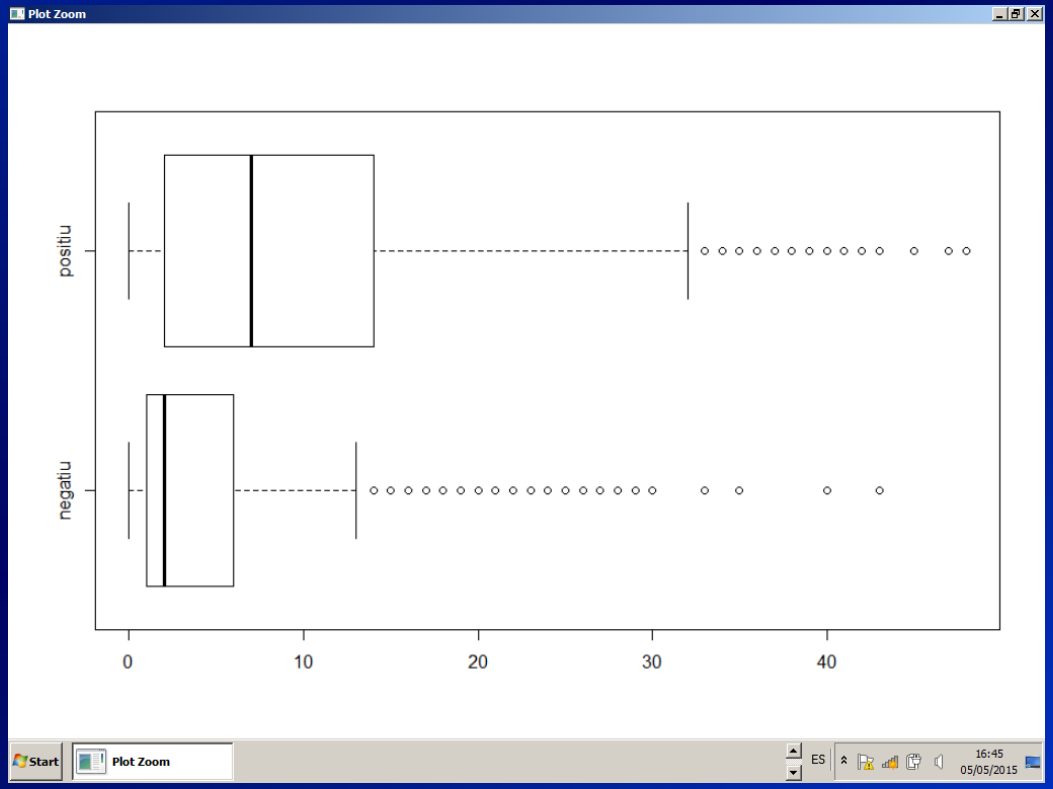
or

Apparently dependent

One categorical variable and one numerical

Multiple boxplot

Antiguitat a la feina



One categorical and one numerical

Descriptive by groups

aggregate(Antiguedad.Trabajo, by=list(Dictamen), FUN=mean)

Group.1 x

1 negatiu 4.586922

2 positiu 9.319062

aggregate(Antiguedad.Trabajo, by=list(Dictamen), FUN=sd)

Group.1 x

1 negatiu 6.118022

2 positiu 8.487919

aggregate(Antiguedad.Trabajo, by=list(Dictamen), FUN=max)

Group.1 x

1 negatiu 43

2 positiu 48

aggregate(Antiguedad.Trabajo, by=list(Dictamen), FUN=median)

Group.1 x

1 negatiu 2

2 positiu 7

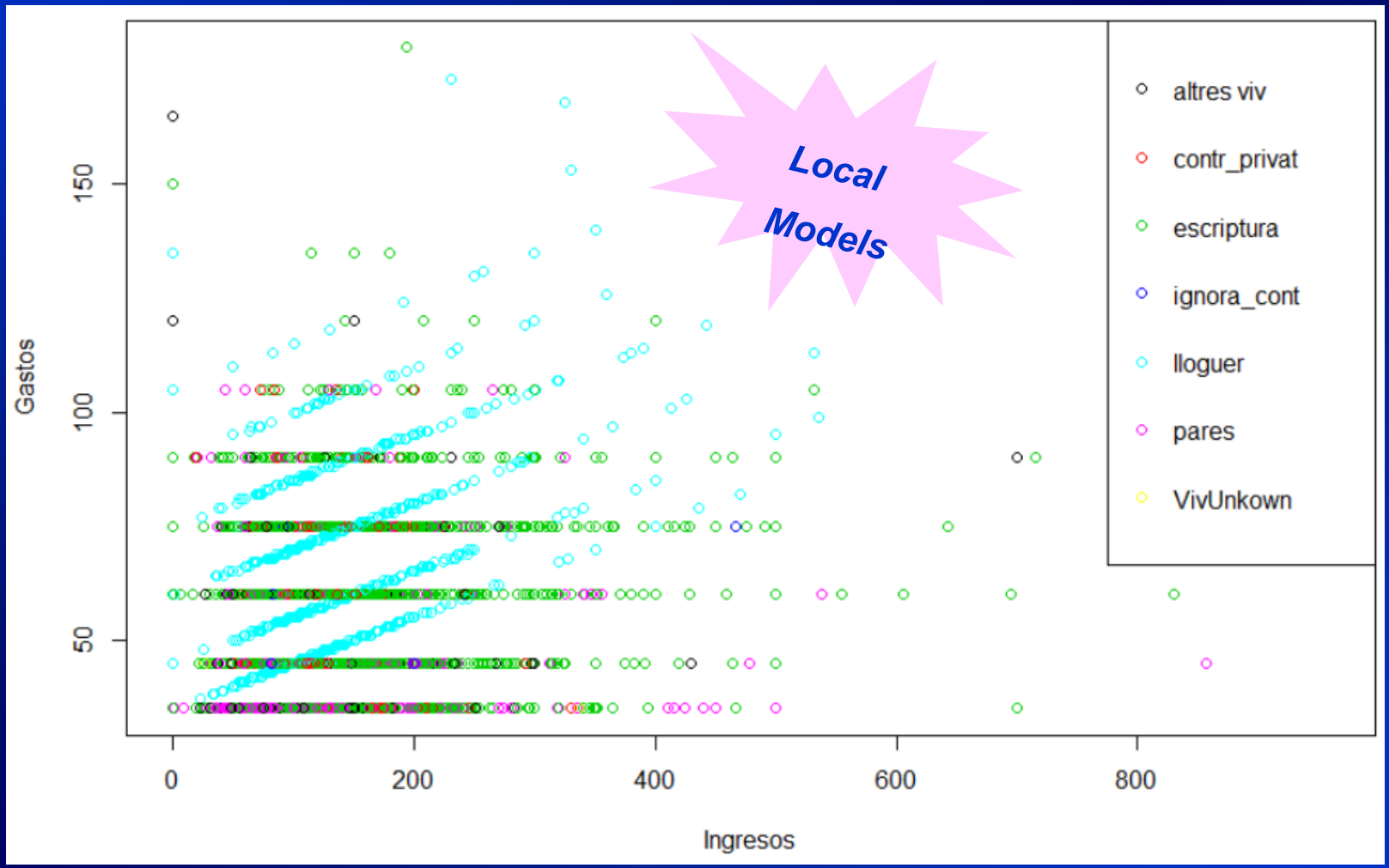
Assessing association between one categorical variable and one numerical

The F Test

Consider Kruskall-Wallis test when required

Two numerical and one categorical

Letter-plot



Bivariate Statistics

Karina Gibert

Dpt. Statistics and Operation Research

Knowledge Engineering and Machine Learning Research group

Universitat Politècnica de Catalunya-BarcelonaTech (Spain)

karina.gibert@upc.edu

www.eio.upc.edu/homepages/karina

Are there any questions?...