



Big Data Analytics Academic Year 2025-2026

Simone Borra

Informations

Week	day	hours	Teaching	
week1	March 5	11:00-13:00	Prof Borra	Introduction
week2	March 12	11:00-13:00	Prof Borra	Introduction
week2	March 13	14:00-17:00	SAS – (Zoom)	SAS Visual Analytics
week2	day1	4 hours	SAS e-learn	SAS Visual Analytics
week3	March 20	14:00-17:00	SAS – (Zoom)	SAS Visual Analytics
week3	Day1	4 hours	SAS e-learn	SAS Visual Analytics
week4	March 27	14:00-17:00	SAS – (Zoom)	SAS Visual Analytics
week4	day1	4 hours	SAS e-learn	SAS Visual Analytics
week4	day2	3 hours	SAS e-learn	SAS Visual Analytics
week8	April 10	9:00-13:00 14:00-16:00	SAS – (Zoom)	SAS Case Study
Final Exam	April 28	3 hours 11:00-14:00	Prof Borra (exam)	



Big Data Analytics Academic Year 2022-2023

Simone Borra

Exam

The final exam consists of two parts:

Part 1 (Theory): Multiple choice test concerning the theoretical aspects
1 hour

Part 2 (Practice): Questions about using SAS Visual Analytics on real data
1 hour

The exam is online, so you must use your own computer



Big Data Analytics Academic Year 2020-2021

Simone Borra

Aims

The course provides an introduction to data preparation, data analysis and report creation in **SAS Visual Analytics**.

Students will learn how to use this point-and-click SAS environment to:

- **access, transform and modify data**
- **visually explore data to discover new insights.**

This Data Visualization tool by SAS enable to easily search for:

- **Relationships**
- **Trends**
- **patterns**



Big Data Analytics Academic Year 2020-2021

Simone Borra

Teaching

Teaching Method

Classroom teaching: a SAS expert will describe the main topics of the course and will answer students' questions.

E-learning: a collection of **videos**, **demos**, and practices, that summarize the concepts shown in classroom.

Case studies: in which students can practice with the supervision of the teacher.

The final exam consists of a written test containing open-ended and/or closed-ended questions, covering all the topics of the course.



Big Data Analytics Academic Year 2020-2021

Simone Borra

Certification

Lectures and e-learning lessons can help students prepare for the certification exam:

SAS Certified Specialist: Visual Business Analyst.

To complete the preparation, the student will have to dedicate time to carry out the e_learning module:

SAS Visual Analytics 2 for SAS Viya: Advanced

All certification details are described in:

<https://www.sas.com/sas/training/scyp.html>.

Roadmap to analyse Big Data



Data
preparation

Describe single
variables

Discover
relationships or
trends or
patterns

Sample data could be collected in several ways:

1. variables are measured on n different units, in a certain instant
2. variables are measured in T different times
3. variables are measured in T different times always on the same n units

consequently we have the following approaches:

1. Cross section analysis
2. Analysis for time series
3. Analysis for panel data

Data preparation

	t=1	
ID	X_1	X_2
12	X_{11}	X_{12}
34	X_{21}	X_{22}
...	...	
25	X_{n1}	X_{n2}

	t=2	
ID	X_1	X_2
33	X_{11}	X_{12}
72	X_{21}	X_{22}
...	...	
86	X_{n1}	X_{n2}

	t=3	
ID	X_1	X_2
43	X_{11}	X_{12}
101	X_{21}	X_{22}
...	...	
122	X_{n1}	X_{n2}

	t=4	
ID	X_1	X_2
6	X_{11}	X_{12}
210	X_{21}	X_{22}
...	...	
66	X_{n1}	X_{n2}

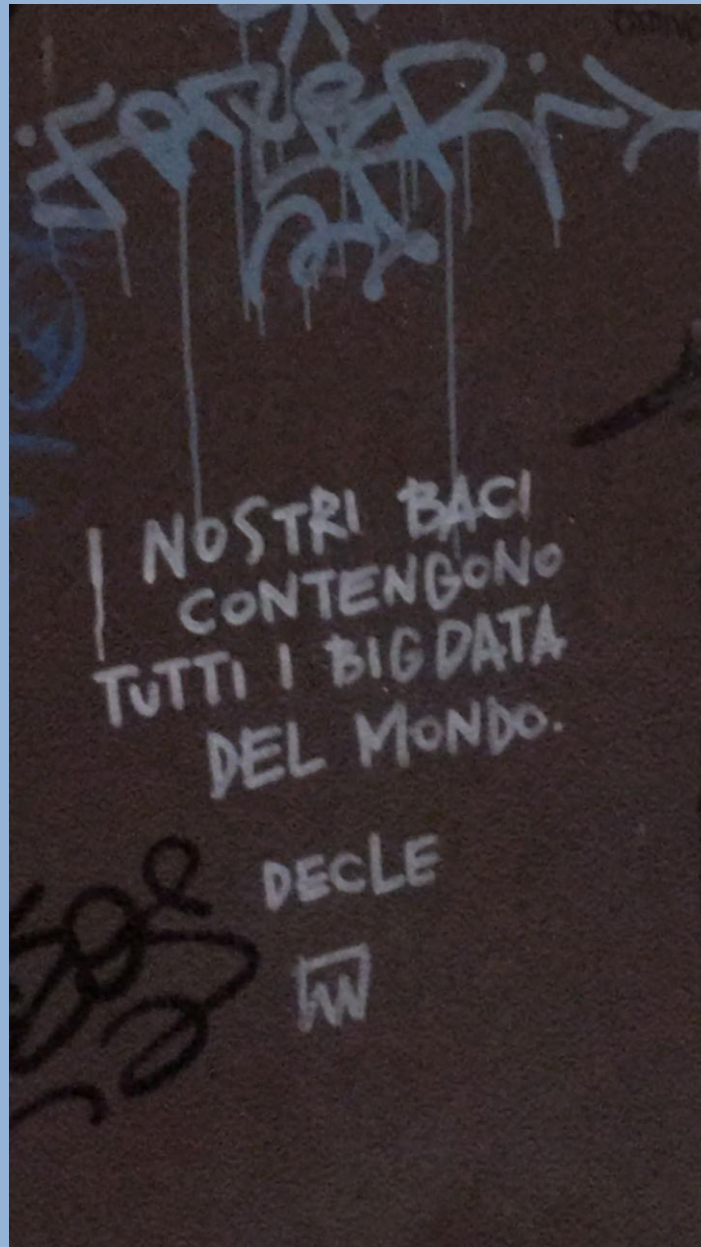
	t=1	
ID	X_1	X_2
12	X_{11}	X_{12}
34	X_{21}	X_{22}
...	...	
25	X_{n1}	X_{n2}

	t=2	
ID	X_1	X_2
12	X_{11}	X_{12}
34	X_{21}	X_{22}
...	...	
25	X_{n1}	X_{n2}

	t=3	
ID	X_1	X_2
12	X_{11}	X_{12}
34	X_{21}	X_{22}
...	...	
25	X_{n1}	X_{n2}

	t=4	
ID	X_1	X_2
12	X_{11}	X_{12}
34	X_{21}	X_{22}
...	...	
25	X_{n1}	X_{n2}

*Our kisses
contain all the
big data in the
world*



Big Data - three Vs definition: **V**olume **V**elocity **V**ariety



Sources of Big Data (UNECE, 2013)

1. Social Networks (*human-sourced information*)

Social networks (Facebook, Twitter,...), Blogs, Pictures (Instagram, Picasa,...) Video (Youtube,...), Internet searchers, Mobile data content (Whatsapp,...), Email,..., ecc.

2. Traditional Business systems (*process-mediated data*)

Data produced by Public Agencies, Data produced by businesses: commercial transactions, Banking/stock records, Credit cards, E-commerce,...ecc.

3. Internet of Things (*machine-generated data*)

Data from fixed sensors:

Home automation; Weather/pollution sensors; Traffic sensors/webcam; Security/surveillance videos/images; Scientific sensors

Data from Mobile sensors (tracking):

Mobile phone location; Cars; Satellite images

Data from computer systems:

Logs; Web logs; ecc

Big Data - Advantages

- With a big number of cases should be able to discover and estimate more complex relationship between variables.
- If we do not have a small sample but a very large number of observations chosen randomly, it is of little relevance to consider the **sampling variability of the estimators** of the population parameters.
- In fact, the sample variability tends to decrease as the sample size increases and we can assume that a very large unbiased sample represents the phenomena and their relationships as occurs in the population.

In general, a large amount of data is not a sufficient condition for obtaining good statistical analyzes

Big Data - Disadvantages

- Big data **often cannot be considered as a random sample** and this could strongly distort the analysis;
- Data are often observed and recorded without taking into account subsequent statistical analyzes, and it is quite common that **some essential variables are missing**.
- Data often comes from different sources and always requires cleaning work that considers missing data and inconsistencies.
- As the number of variables increases the possibility of **spurious associations between variables**.

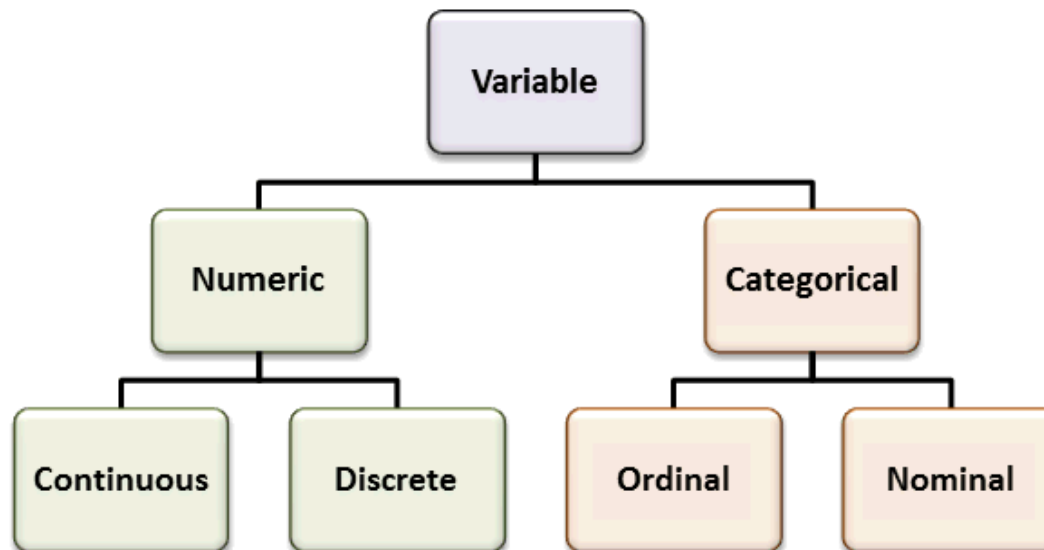
Data type

For the first time in history we have data everywhere, the now called **Big Data**. These data are a mixture of **structured** and **unstructured** information.

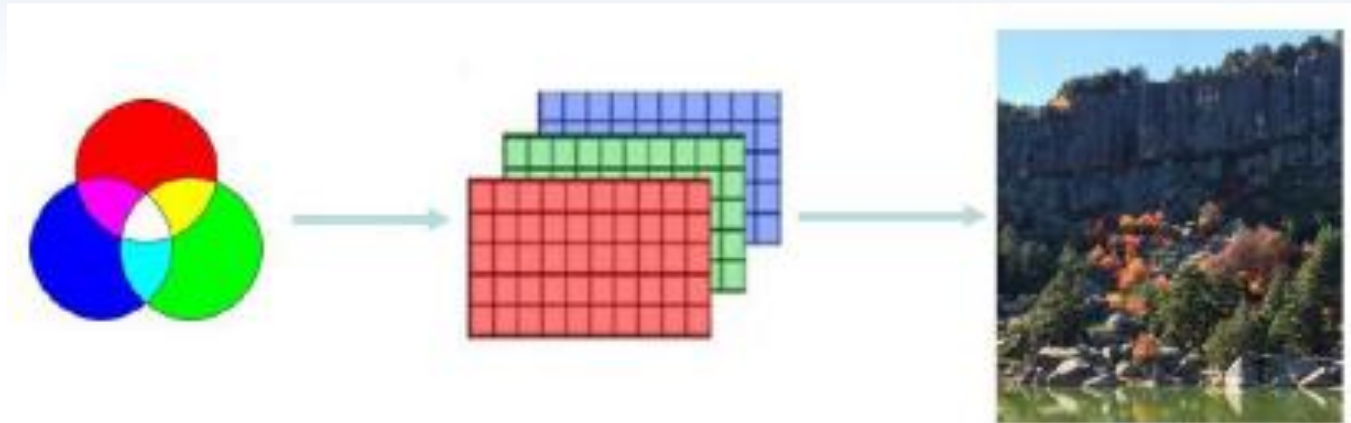
P.Galeano-D. Pena (2019)

Large heterogeneous databases sometimes **unstructured**, which may include **texts**, **images**, **videos**, or **sounds**, from different populations and as many (or even more) variables than observations.

We will only consider **structured data**:



Example: Process of transforming pixels in images



For example, a grid $10 \times 10 = 100$ cells

In each cell we have three numbers:

Intensity of red; intensity of green; intensity of Blue

We have $100 \times 3 = 300$ numbers to transform a set of pixels into a image

Data type

Depending on the type of variable we can use different statistical tools to reclassify, transform and analyze variables.

Data preparation

Using several sources of data we need to homogenize data and integrate different tables of data.

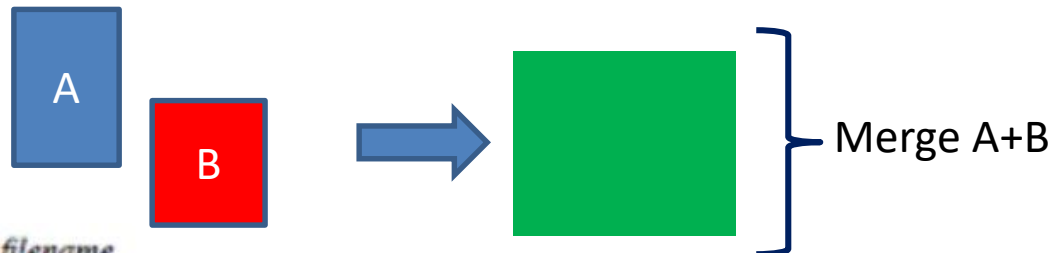
Data preparation

Using several sources of data we need to homogenize data and integrate different tables of data.

Merge

Merge adds columns, or variables. You would use merge when there are the same observations across datasets with different variables.

Many-to-one-merge



```
. merge m:1 region using filename
      master      +      using      =
```

id	region	a
1	2	26
2	1	29
3	2	22
4	3	21
5	1	24
6	5	20

region	x
1	15
2	13
3	12
4	11

=

merged result

id	region	a	x	_merge
1	2	26	13	3
2	1	29	15	3
3	2	22	13	3
4	3	21	12	3
5	1	24	15	3
6	5	20	.	1
.	4	.	11	2

One-to-Many-merge

```
. merge 1:m region using filename
```

master

+

using

=

merged result

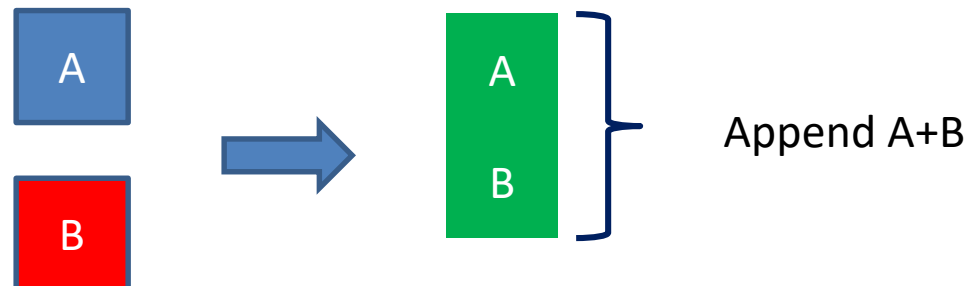
region	x
1	15
2	13
3	12
4	11

id	region	a
1	2	26
2	1	29
3	2	22
4	3	21
5	1	24
6	5	20

region	x	id	a	_merge
1	15	2	29	3
1	15	5	24	3
2	13	1	26	3
2	13	3	22	3
3	12	4	21	3
4	11	.	.	1
5	.	6	20	2

Append

Is used when you want combine datasets that contain the same variables, but have different cases, thus, you are adding rows to the dataset, but the number of columns should remain the same.



Data integration

Set of Techniques for joining different tables and ensuring data consistency: measurement units, type of variable ...

customer	Age	Educ level
Paolo	23	Secondary school dipl
Francesca	45	Degree
Erik	41	Secondary school dipl
Maria	29	Primary school dipl

Data-set A

customer	Age	Educ level
Erik	41	Medium
John	27	High
Maria		Low
Alan	41	Low

Data-set B

customer	Age	Educ level
Erik	41	Medium
John	27	High
Maria	29	Low
Paolo	23	Medium
Francesca	45	High
Alan	41	Low

Data-set A+B

Educational level "Primary school dipl." = "Low"

Educational level "Secondary school dipl." = "Medium"

Educational level "Degree" = "High"

Transformation of **nominal** variables into **quantitative** variables

We use **dummy** variables, with only two values: 1 or 0.

Example: **Gender** Male=0 Female=1

If the variable has K modalities we use K-1 dummy variables.

For instance, K=3 we use 2 dummies:

Name	Employment status	Dummy1	Dummy2
Jack	Full-time employed	1	0
Anna	Part-time employed	0	1
Mark	Unemployed	0	0

Transformation of **ordinal** variables into **quantitative** variables

For each modality we use an interger value

Example:

Educational level

Elementary=1 Middle=2 High=3 Degree=4 Post-Degree=5

Degree of satisfaction

Very dissatisfied=-2

Dissatisfied =-1

Neutral=0

Satisfied=+1

Very satisfied=+2

The reclassification is a very discretionary that could greatly influence subsequent analysis

Subdivision of the quantitative variable into classes

In many situations it is convenient to divide the quantitative character into different classes of values

Example:

In a Labour forces context these classes could be used

Age in 3 classes :

from 0 to 16; from 16 to 65; greater than 65 up to 110.

student

employed

retired

- The number of classes must be adequate to the problem
- Classes must not overlap
- Classes must include all possible values

For many statistical analysis it is convenient to first treat the characters in order to make them homogeneous with each other.
In particular, two quantitative characters can differ for:

- **measurement unit** (weight in Kg, height in cm)
- **different intensity** (average weight between adults is 70 Kg, average weight between newborns is 3,5 Kg)
- **different variability** (the income could vary between 500 and 3000 euro, or between 500 and 10.000 euro)

In order to homogenize the observed characters, **normalization** and **standardization** techniques can be applied to them.

How to transform the values of a feature in such a way that all the transformed values are within a certain range.

For instance, we observe the following values for X:

2; 3; 4; 12; 23; 32; 34; 67

The minimum and maximum values are min=2 e max=67.

We want to transform the values so that the new feature X' has min_new = 1 and max_new = 10.

We apply on the original values this function:

$$X' = \frac{X - \min}{\max - \min} \times (\max_new - \min_new) + \min_new$$

If X=2 we obtain:

$$X' = \frac{2 - 2}{67 - 2} \times (10 - 1) + 1 = 0 + 1 = 1$$

If X=67 we obtain:

$$X' = \frac{67 - 2}{67 - 2} \times (10 - 1) + 1 = 9 + 1 = 10$$

The values of X' are:

1; 1,39; 2,38; 3,91; 5,15; 5,43; 6,82 ; 10

How to transform the values of a feature so that the mean will be 0.

For instance, we observe the following values for X:

2; 3; 12; 23; 32; 34; 44; 67

The mean is $(12+3+32+67+23+44+2+34)/8=27,125$

To transform the values of a feature so that the mean will be 0, we apply:

$$X' = X - \text{mean}(X)$$

Obtaining:

-25,125; -24,125; -15,185; -4,125; 4,875; 6,875; 16,875; 39,875

and now the average value is equal to 0.

STANDARDIZATION: How to transform the values of a feature so that the mean will be 0 and variance 1.

For instance, we observe the following values for X:

2; 3; 12; 23; 32; 34; 44; 67

The mean is $(2+3+12+23+32+34+44+67)/8=27.125$

The variance is $[(2-27.125)^2+\dots+(67-27.125)^2]/8=425.6$

The standard deviation $=\sqrt{425.6} = 20.6$

To transform the values of a feature so that the mean will be 0 variance 1, we apply:

$$X' = \frac{X - \text{mean}(X)}{\text{stand.dev.}(X)}$$

Obtaining:

-1.218; -1.169; -0.733; -0.200; 0.236; 0.333; 0.818; 1.933

and now the average value is equal to 0 and the variance is 1.

The data may contain **missing values**, i.e. for some statistical units only a part of the variables of interest may have been detected.

For instance, Bianchi's age is missing, the Dotti's is missing.

Name	Age	Sex	Educ lev	Activity	Weight (kg)	Exam Score
Rossi M.	32	M	degree	employed	72	65
Bianchi G.		F	degree	employed	55	55
Nicoletti C.	46	M	diploma	unemployed	79	53
Marcelli F.	28	M	diploma	student	63	78
Petrone A.	51	F	diploma	housewife	64	21
Dotti P.	33	M	degree		64	66

Data preparation

There are several techniques of dealing with missing data.

The simplest, but also the most expensive in terms of information loss, is to eliminate cases corresponding to the missing values.

Name	Age	Sex	Educ lev	Activity	Weight (kg)	Exam Score
Rossi M.	32	M	degree	employed	72	65
Bianchi G.		F	degree	employed	55	55
Nicoletti C.	46	M	diploma	unemployed	79	53
Marcelli F.	28	M	diploma	student	63	78
Petrone A.	51	F	diploma	housewife	64	21
Dotti P.	33	M	degree		64	66

this method leads to a reduction in the sample size. Consequences: **biased estimates, reduced reliability of estimates.**

Data preparation

If the variable is **quantitative**, another method is to replace the missing value with the average (median) of the observed values on the remaining statistical units.

For instance, The mean value of Age is:

$$(32+46+28+51+33)/5=38$$

Name	Age	Sex	Educ lev	Activity	Weight (kg)	Exam Score
Rossi M.	32	M	degree	employed	72	65
Bianchi G.	38	F	degree	employed	55	55
Nicoletti C.	46	M	diploma	unemployed	79	53
Marcelli F.	28	M	diploma	student	63	78
Petrone A.	51	F	diploma	housewife	64	21
Dotti P.	33	M	degree		64	66

Consequence: this method **reduce the feature variability**

Another method is to replace the missing value with that observed statistical unit most similar to that considered. There are many ways to define the "similarity" between two statistical units.

Dotti P. presents values with respect to age, sex, educational qualification and score very "close" to those presented by Rossi. Therefore, Dotti's activity is attributed to Rossi, that is "employed".

Name	Age	Sex	Educ lev	Activity	Weight (kg)	Exam Score
Rossi M.	32	M	degree	employed	72	65
Bianchi G.	38	F	degree	employed	55	55
Nicoletti C.	46	M	diploma	unemployed	79	53
Marcelli F.	28	M	diploma	student	63	78
Petrone A.	51	F	diploma	housewife	64	21
Dotti P.	33	M	degree	employed	64	66

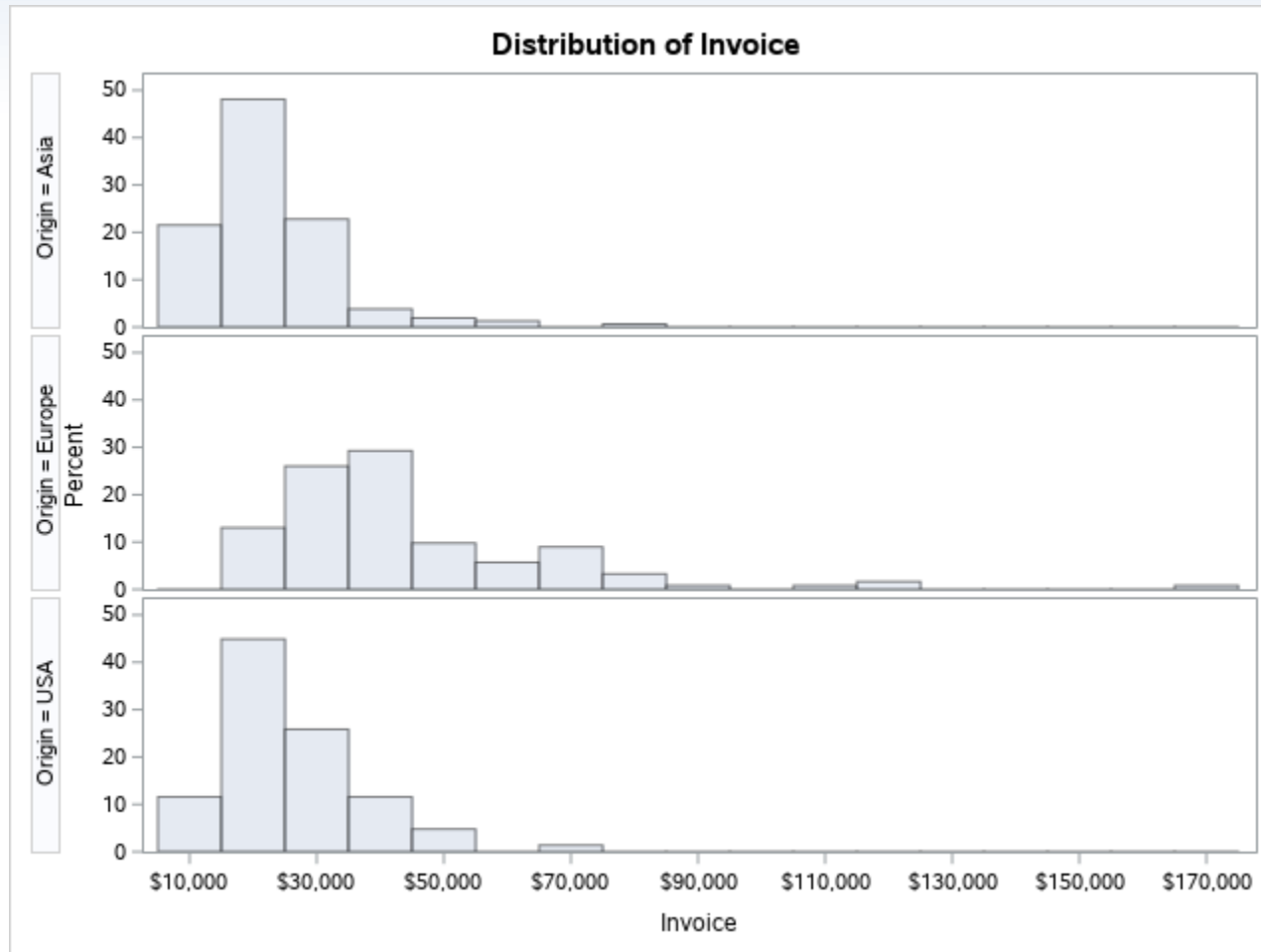
donor imputation is to fill in the missing values for a given unit by copying observed values of another unit, the donor.

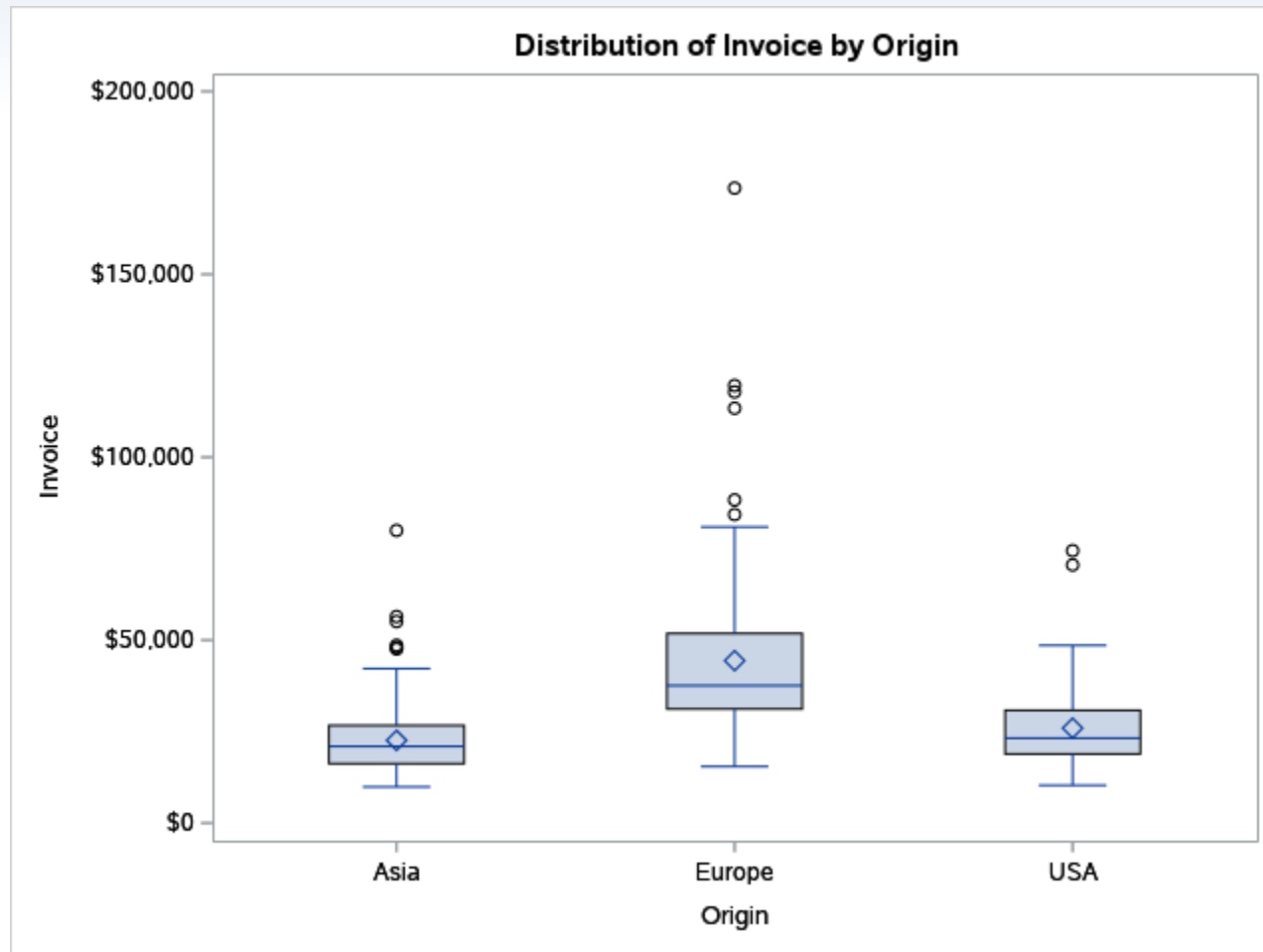
Description of a single feature

For a categorical/quantitative variable

- Frequency distribution
- Position Indices
- Dispersion Indices
- % of missing data
- Presence of outliers or extreme values

Feature description





Feature description

Analysis Variable : Invoice

Origin	N Obs	Mean	Std Dev	Minimum	Maximum	Median	N
Asia	158	22602.18	9842.98	9875.00	79978.00	20949.50	158
Europe	123	44395.08	23080.37	15437.00	173560.00	37575.00	123
USA	147	25949.34	10518.72	10319.00	74451.00	23217.00	147

Association

Several indices in order to measure association between two variables:

- Nominal vs Nominal: Chi-square $(0 ; +\infty)$, V-Cramer $(0 ; 1)$
- Ordinal vs Ordinal: Gamma $(-1 ; +1)$, Tau-b Kendal $(-1 ; +1)$
- Quantitative vs Quantitative: Correlation index $(-1 ; +1)$

Were you satisfied with your experience today?

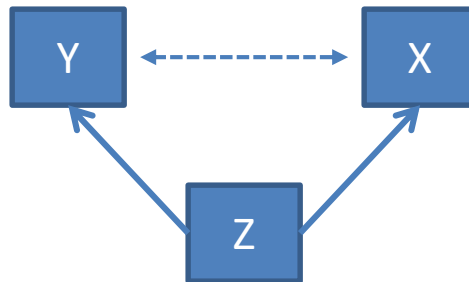
1=very Unsatisfied, 2, 3, 4, 5= very Satisfied

Did our product/service meet your expectations?

1=absolutely No, 2, 3, 4, 5= absolutely Yes

Spurious correlation

In simple linear regression frequently we have spurious correlation, when two variables Y and X have no direct causal connection, yet it may be wrongly inferred that they do, due to either coincidence or the presence of a certain third, unseen factor Z .

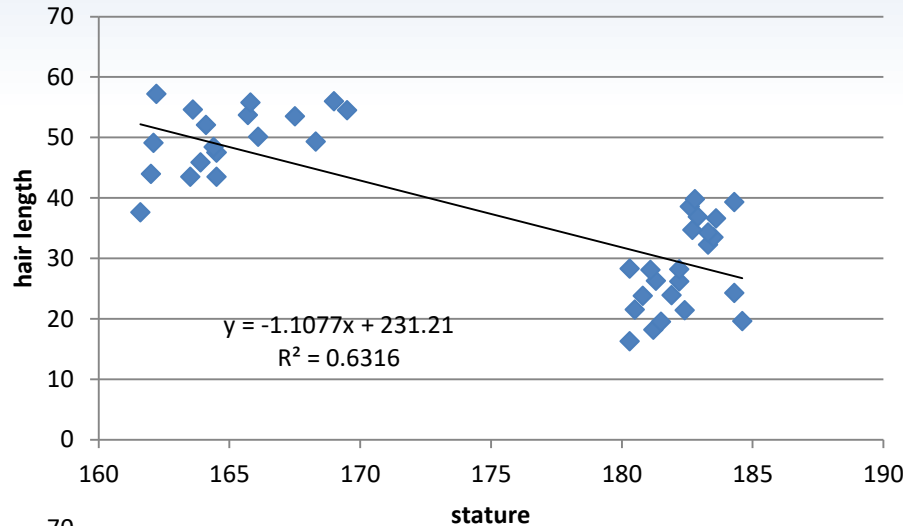


For instance we consider a sample student

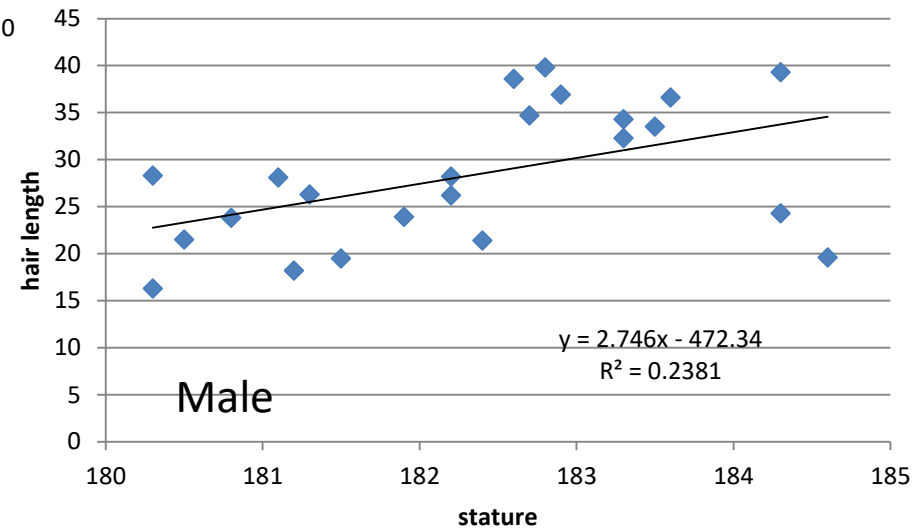
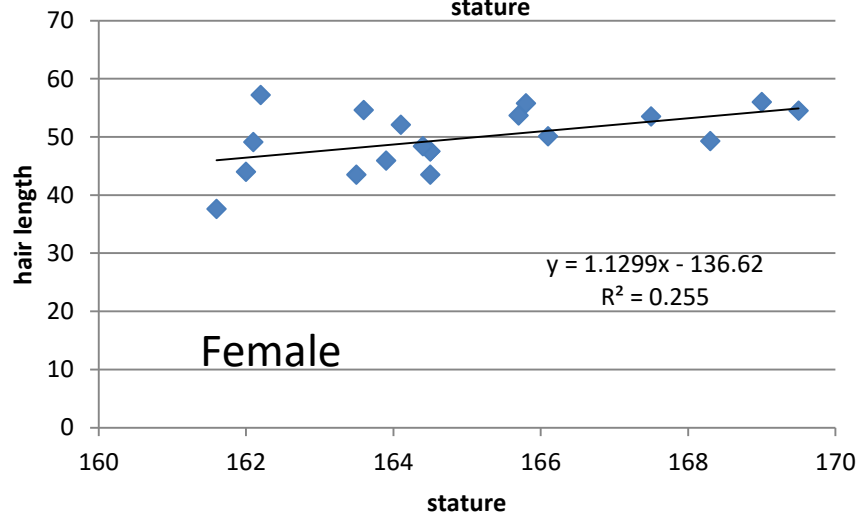
We observe Stature and Hair length

We estimate simple linear regression with Y = Hair length and X =Stature

Spurious correlation



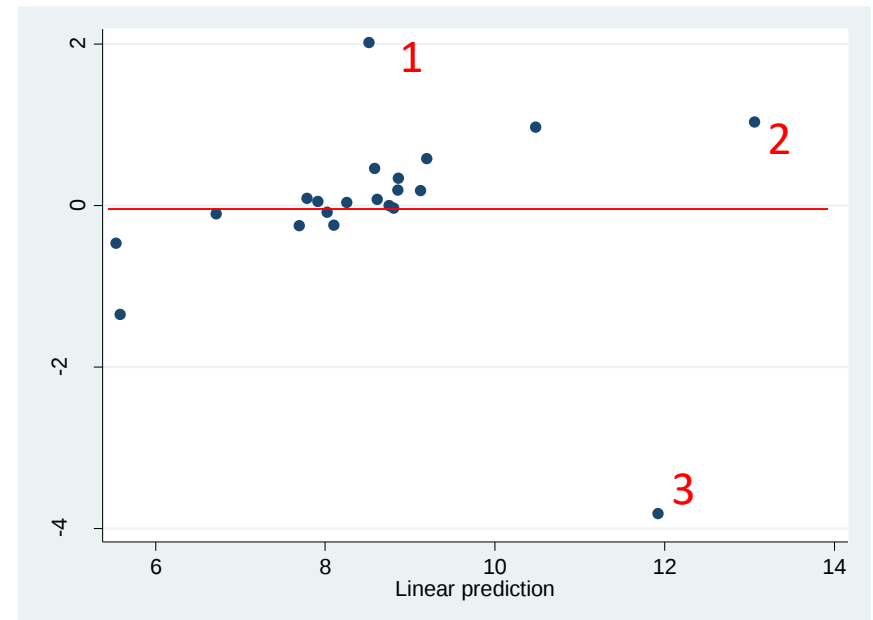
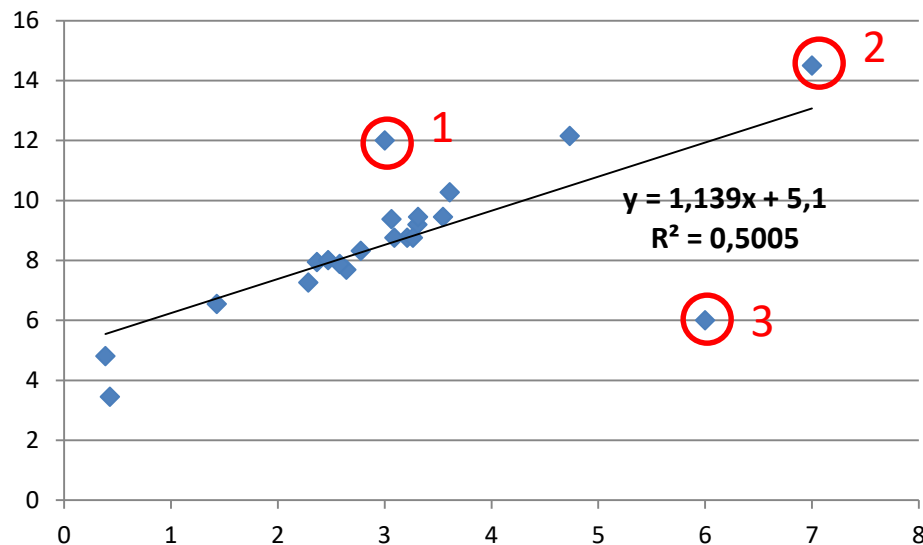
but considering the **sex** we obtain



Presence of outliers

Three different types of outliers:

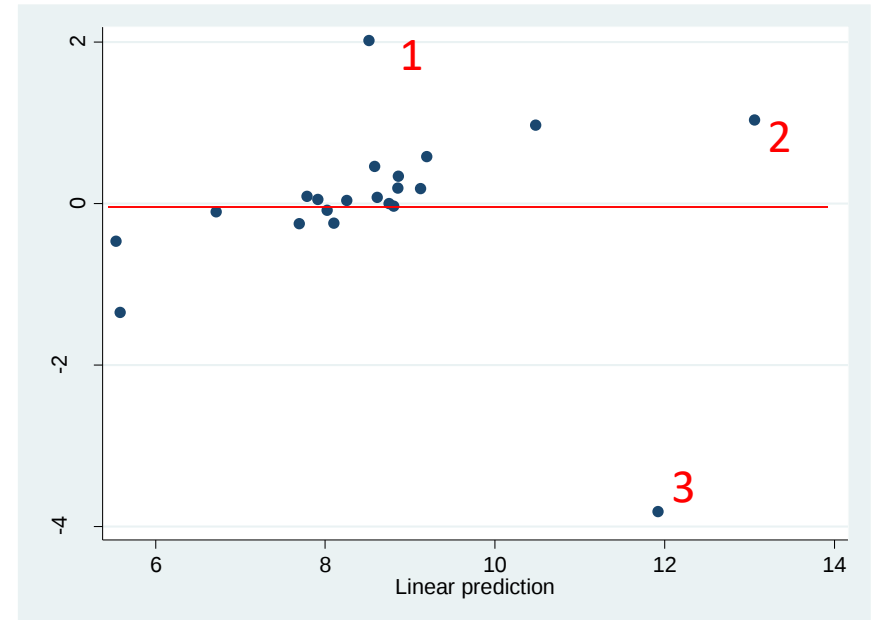
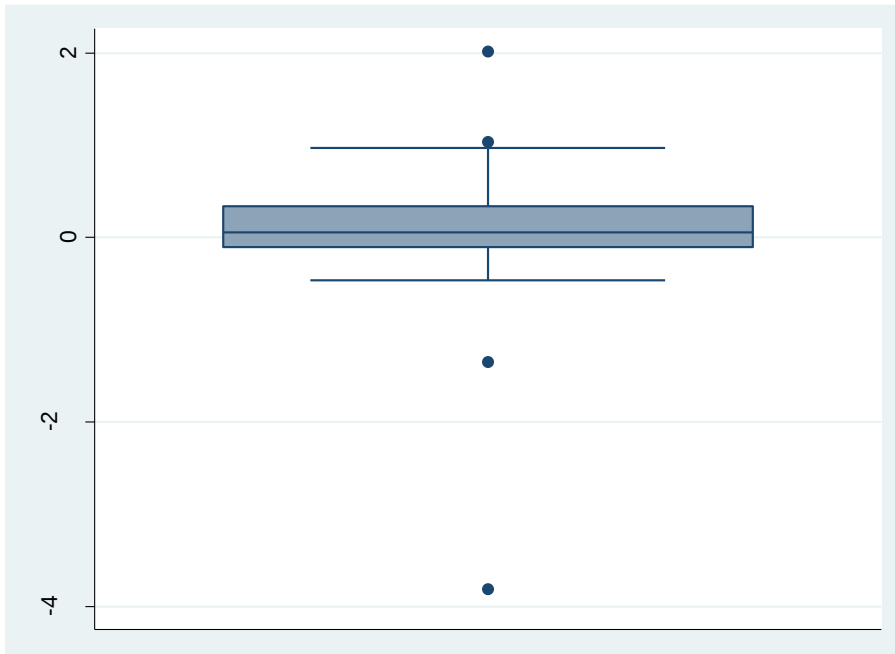
- 1: a value of Y very different from the whole sample but not X
- 2: values of X and Y very different from the whole sample
- 3: a value of X very different from the whole sample but not Y



Presence of outliers

Three different types of outliers:

- 1: a value of Y very different from the whole sample but not X
- 2: values of X and Y very different from the whole sample
- 3: a value of X very different from the whole sample but not Y



Fit lines can be added to scatter plots and heat maps to plot the relationship between variables. The following types of fit lines are available:

Linear – creates a linear fit line (a straight line that best represents the relationship between measures) using a linear regression algorithm.

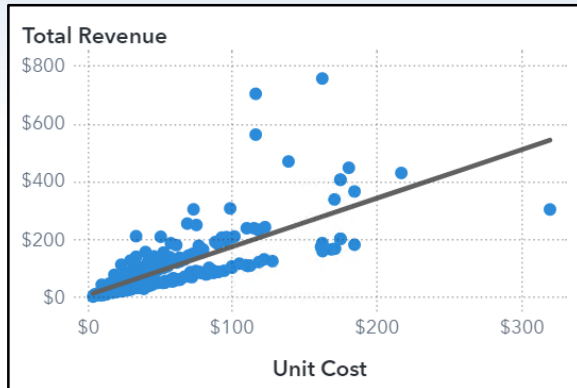
Quadratic – creates a quadratic fit line (a line with a single curve that best represents the relationship between measures).

Cubic – creates a cubic fit line (a line with two curves that best represents the relationship between measures). This method often produces a line with an S shape.

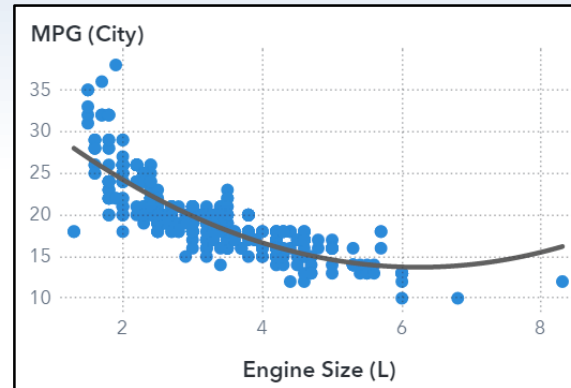
PSpline – creates a penalized B-spline, which is a smoothing spline that closely fits the data. This method can display a complex line with many changes in its curvature.

Fit Line

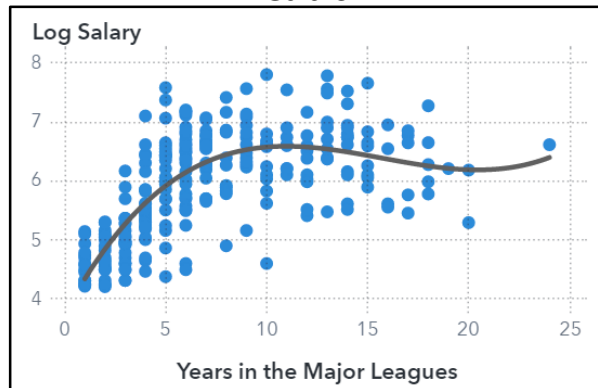
Linear



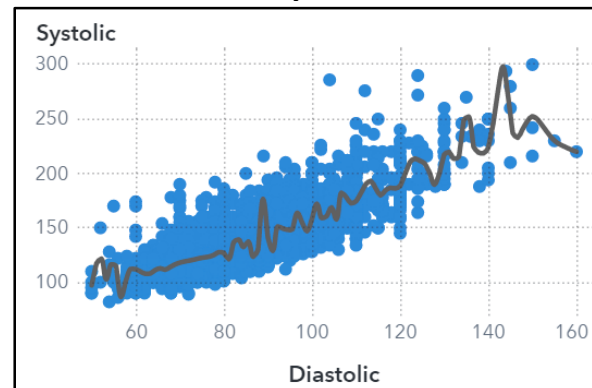
Quadratic



Cubic



PSpline



Running mean smoother

$$s(x_i) = \text{ave}_{j \in I_K(x_i)}(y_j)$$

Running line smoother

Average

we compute a least-squares line instead of a mean in each neighbourhood.

$$s(x) = \hat{\alpha} + \hat{\beta}x \quad x \in I_K(x_i)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are LS estimates for the data points in $I_K(x_i)$.

Locally weighted running line smoother

Average

For each point in $I_K(x_0)$ we give a weight w_i calculated by means of **Tri-cube** function:

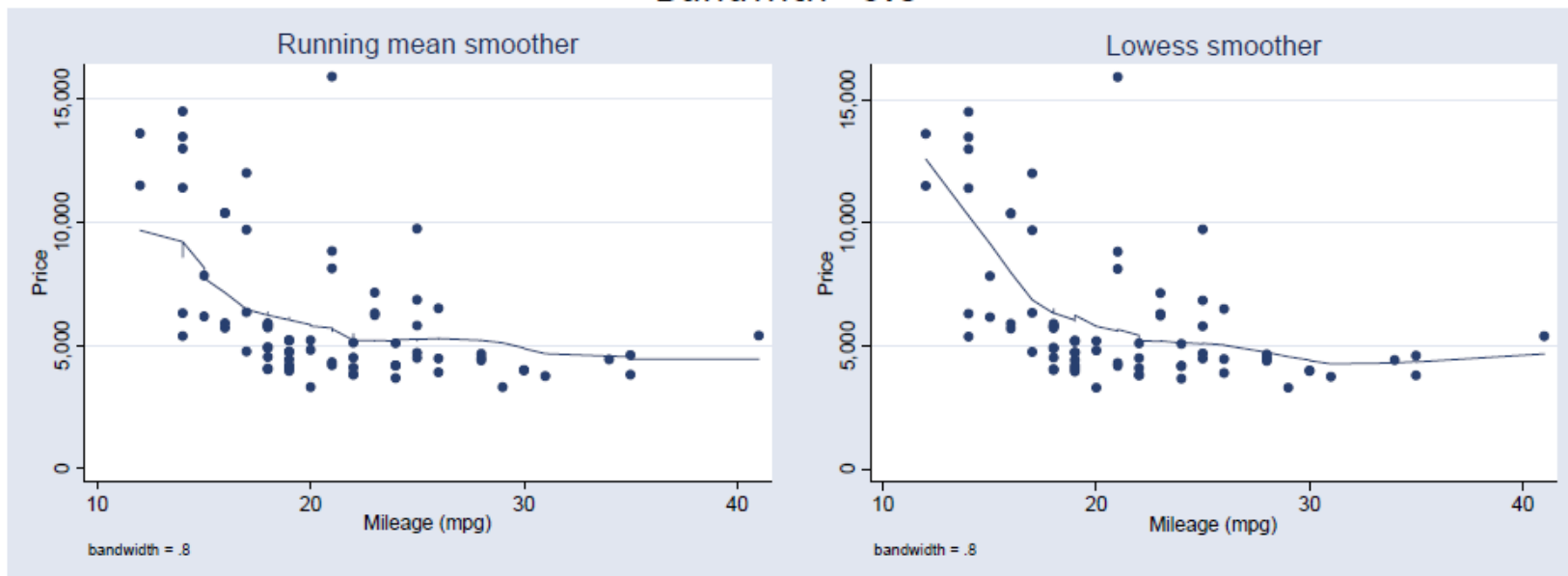
$$W\left(\frac{|x_0 - x_i|}{d_K}\right) \quad W(u) = \begin{cases} \left(1 - |u|^3\right)^3 & |u| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

we compute a weighted least-squares line in each neighbourhood.

$$s(x_0) = \hat{\alpha}(x_0) + \hat{\beta}(x_0)x_0$$

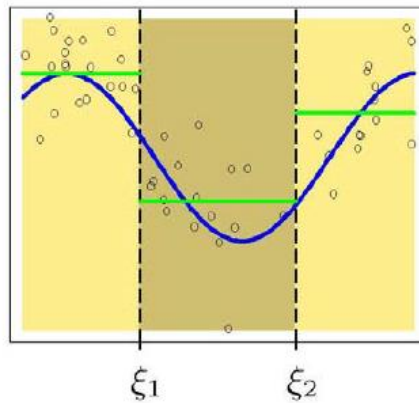
Fit Line

Bandwidth=0.8

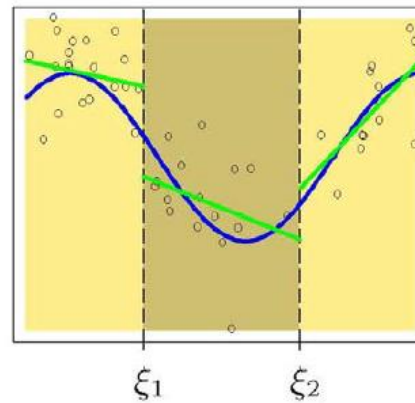


Spline - linear

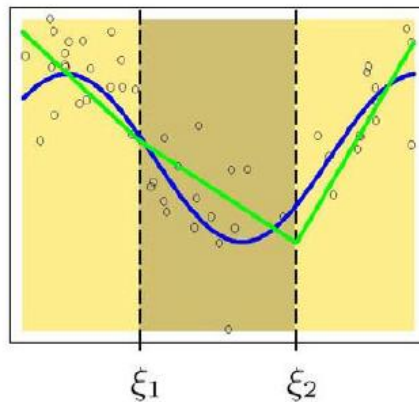
Piecewise Constant



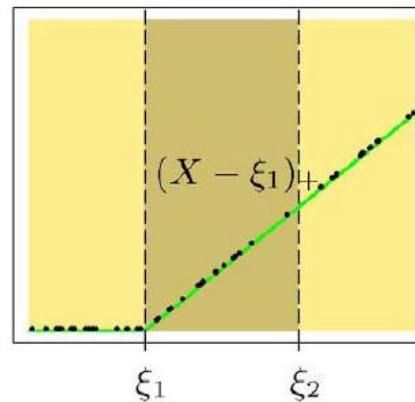
Piecewise Linear



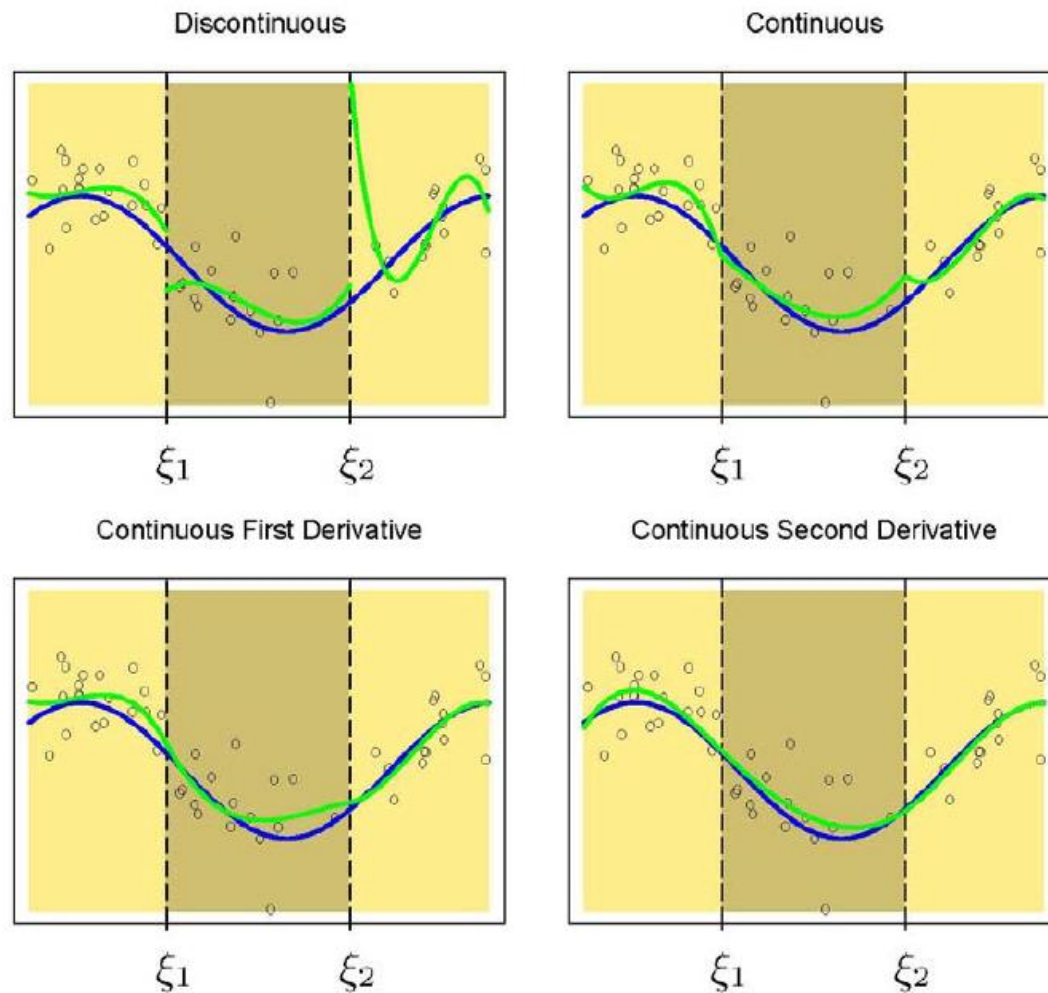
Continuous Piecewise Linear



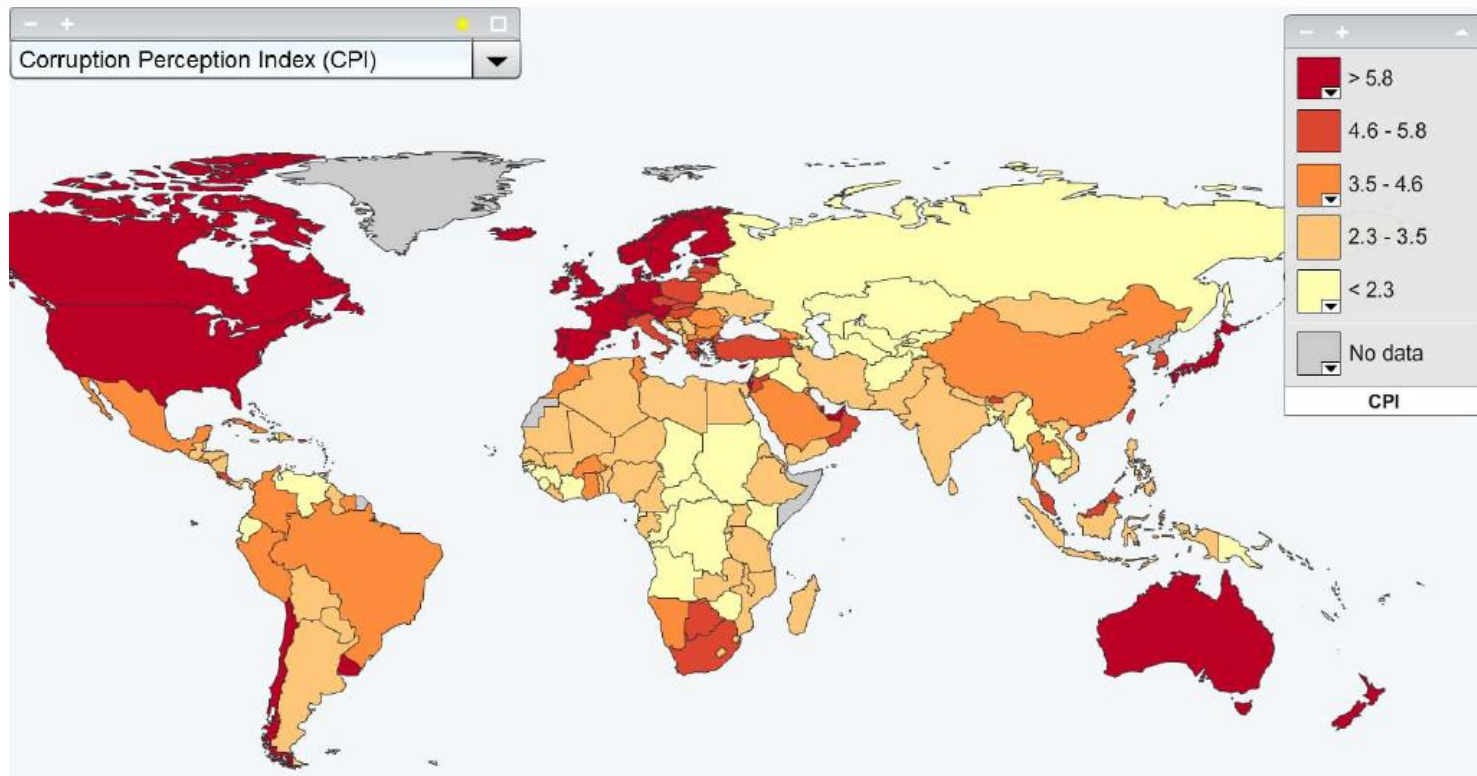
Piecewise-linear Basis Function



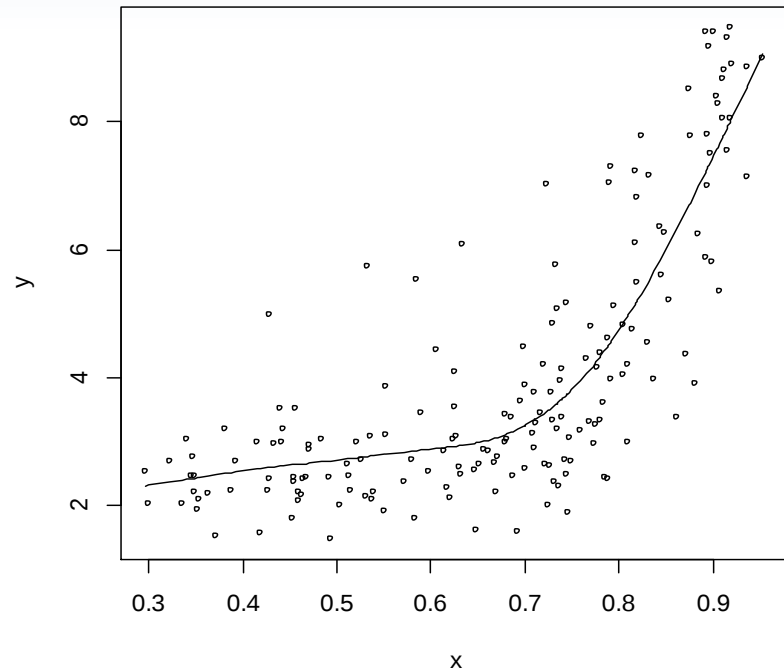
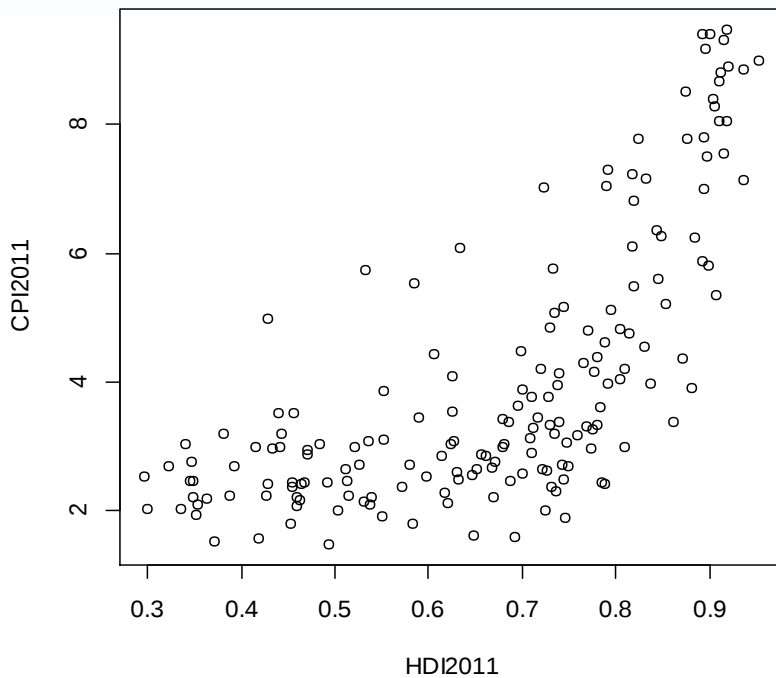
Spline - cubic



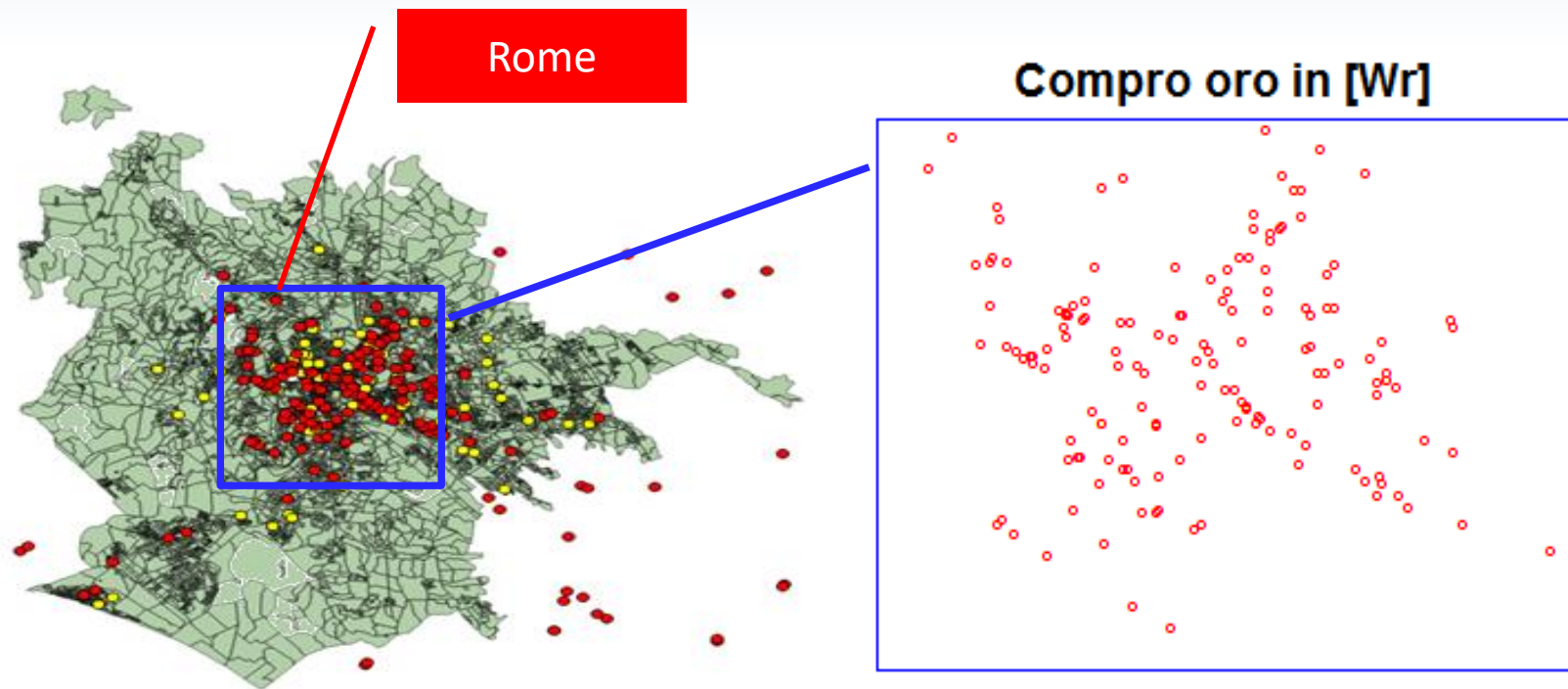
CPI=Corruption perception index (Transparency International)



CPI=Corruption perception index (Transparency International)
HDI=Human Development Index



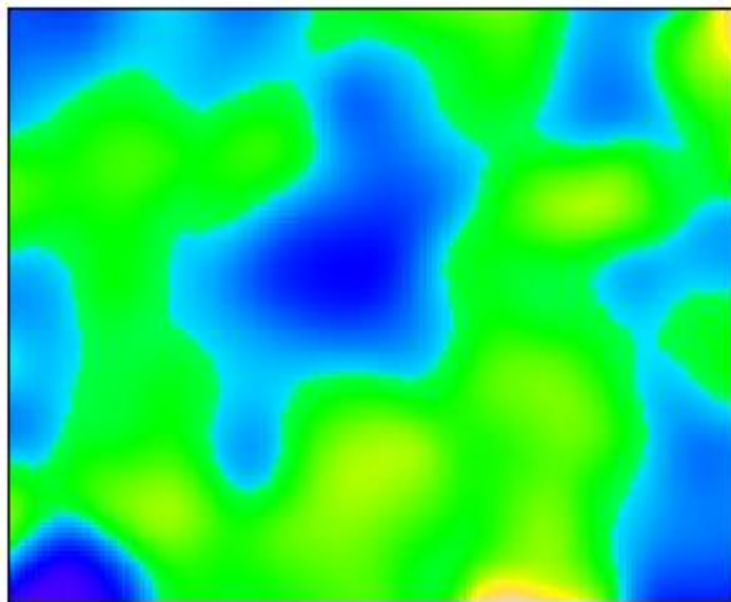
During the economic crisis, "buy gold" shops appeared in many Italian cities.



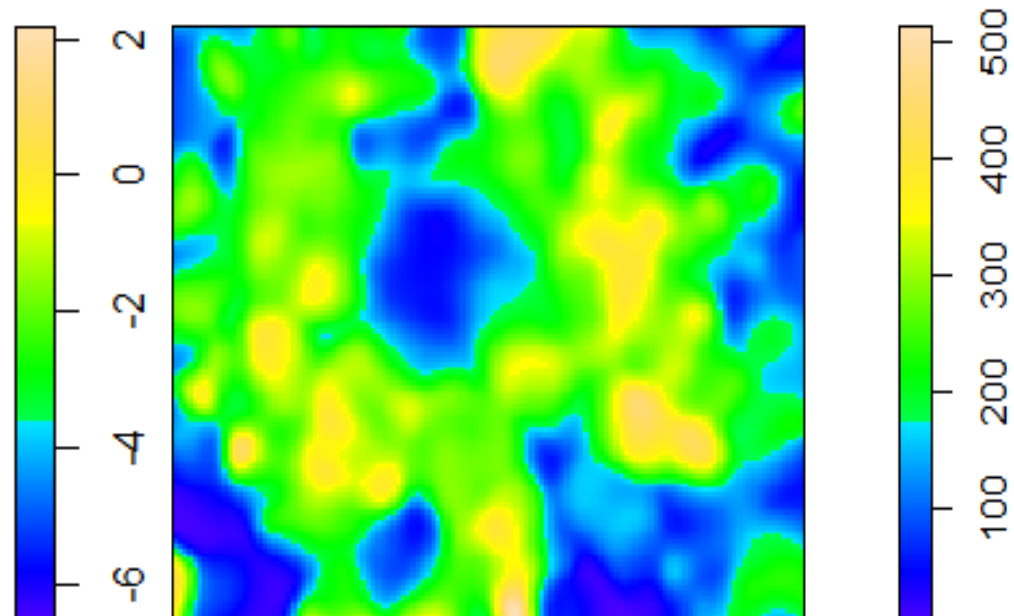
Fit distribution

Distribution of some variables on the territory (Rome) to also have a visual approach to understanding the possible relationships between distribution of points and characteristics of the territory.

Debts distribution



Population distribution



DATA RegHome

Sample size = 67

- **Price**=Home selling price in hundreds of dollars.
- **SQFT**=Square feet of living space
- **Age**=Age of home (years)
- **Feats**=Features Number out of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access)
- **NE**=Located in northeast sector of city (1) or not (0)
- **COR**=Corner location (1) or not (0)

DATA Explore univariate – Stata commands

Descriptive statistics

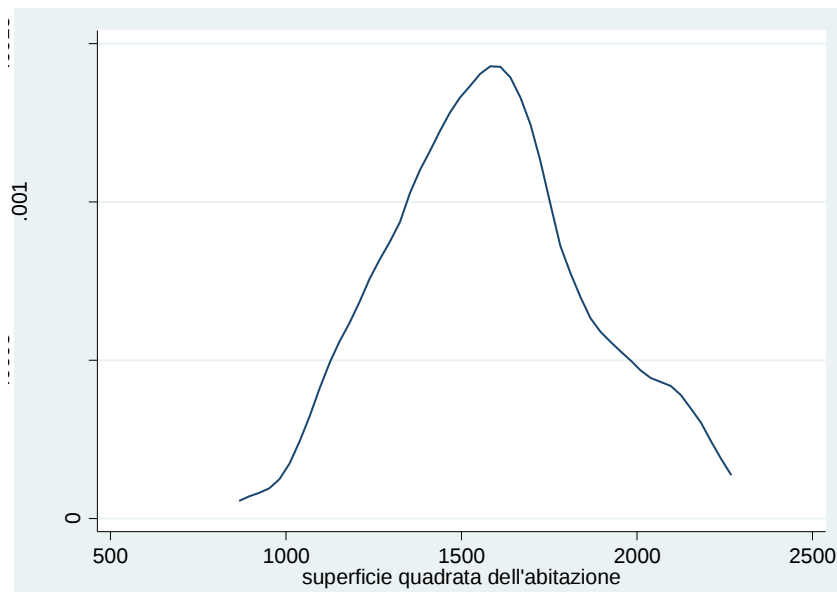
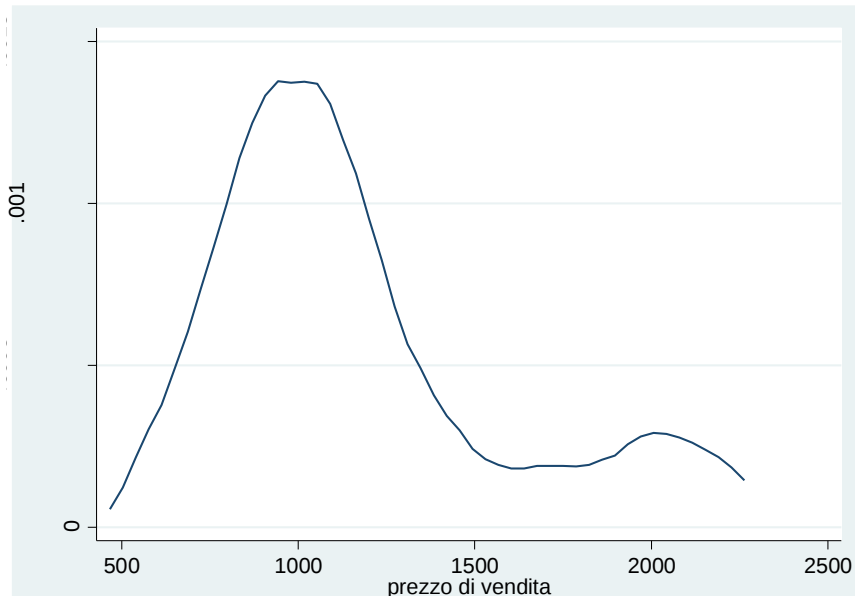
```
. sum price sqrt
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	67	1161.463	405.5376	580	2150
sqft	67	1582.112	278.7009	970	2165.5

DATA Explore univariate – Stata commands

Pdf Estimation

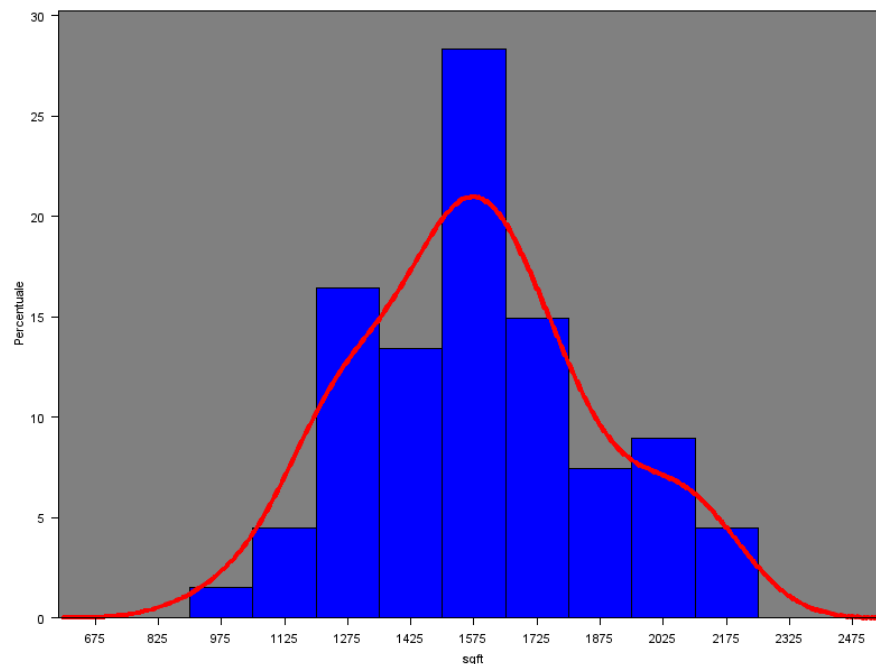
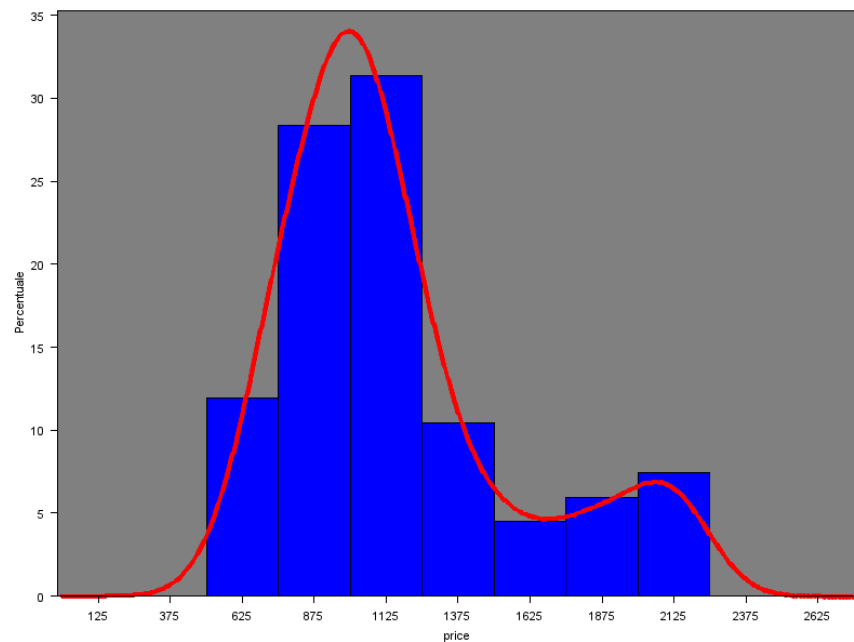
- . kdensity price
- . kdensity sqft



DATA Explore univariate – Stata commands

Pdf Estimation

```
. kdensity price  
. kdensity sqft
```



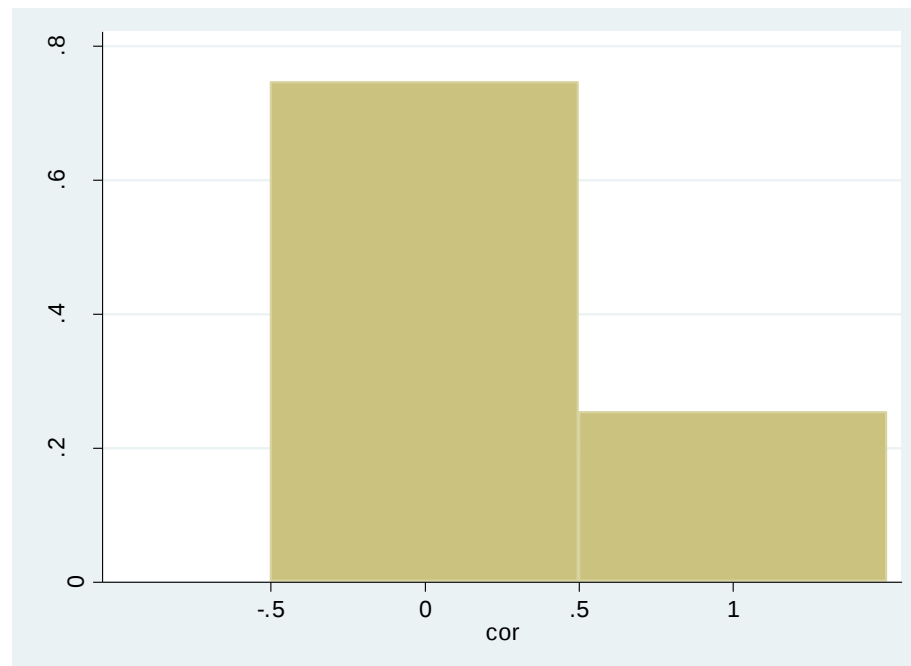
DATA Explore univariate – Stata commands

Descriptive statistics

```
. tab cor
```

collocata			
sull'incroc			
io di due			
strade	Freq.	Percent	Cum.
-----+-----			
0	50	74.63	74.63
1	17	25.37	100.00
-----+-----			
Total	67	100.00	

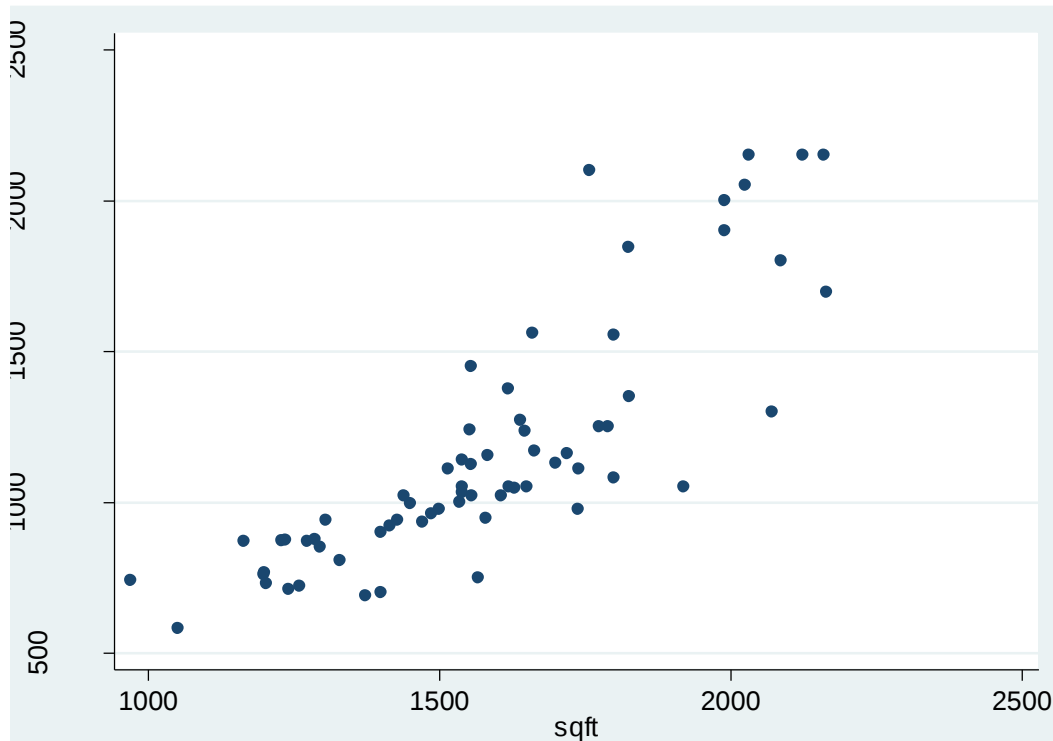
```
. hist cor, discrete
```



DATA Explore Bivariate – Stata commands

Descriptive statistics

- corr price sqft
- scatter price sqft



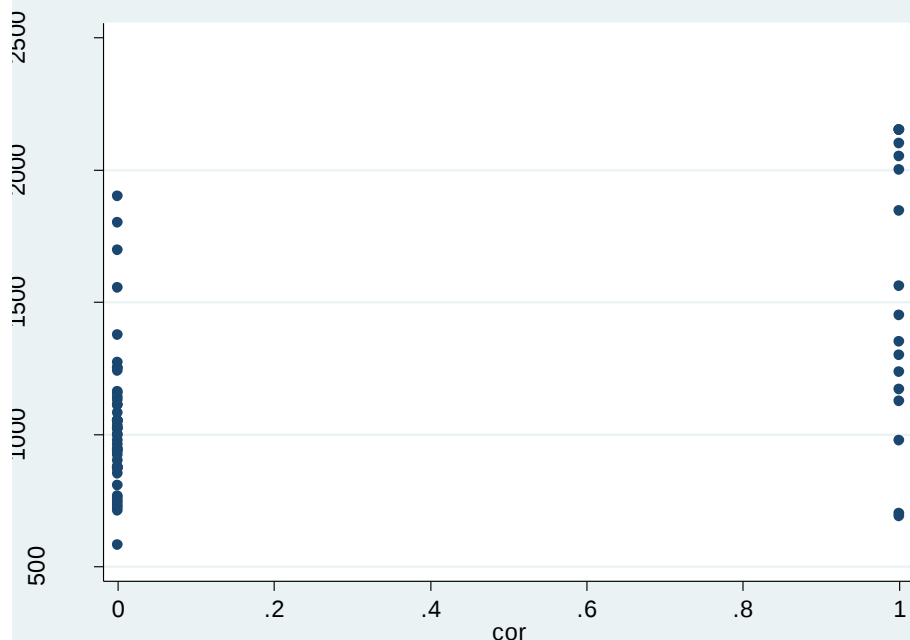
(obs=67)

	price	sqft
price	1.0000	
sqft	0.8341	1.0000

DATA Explore Bivariate – Stata commands

Descriptive statistics

```
. sort cor
. by cor: sum price
. scatter price cor
```



```
-> cor = 0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	50	1036.46	268.4092	580	1900

```
-> cor = 1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	17	1529.118	515.0911	690	2150

If the house is in the corner
the price increases

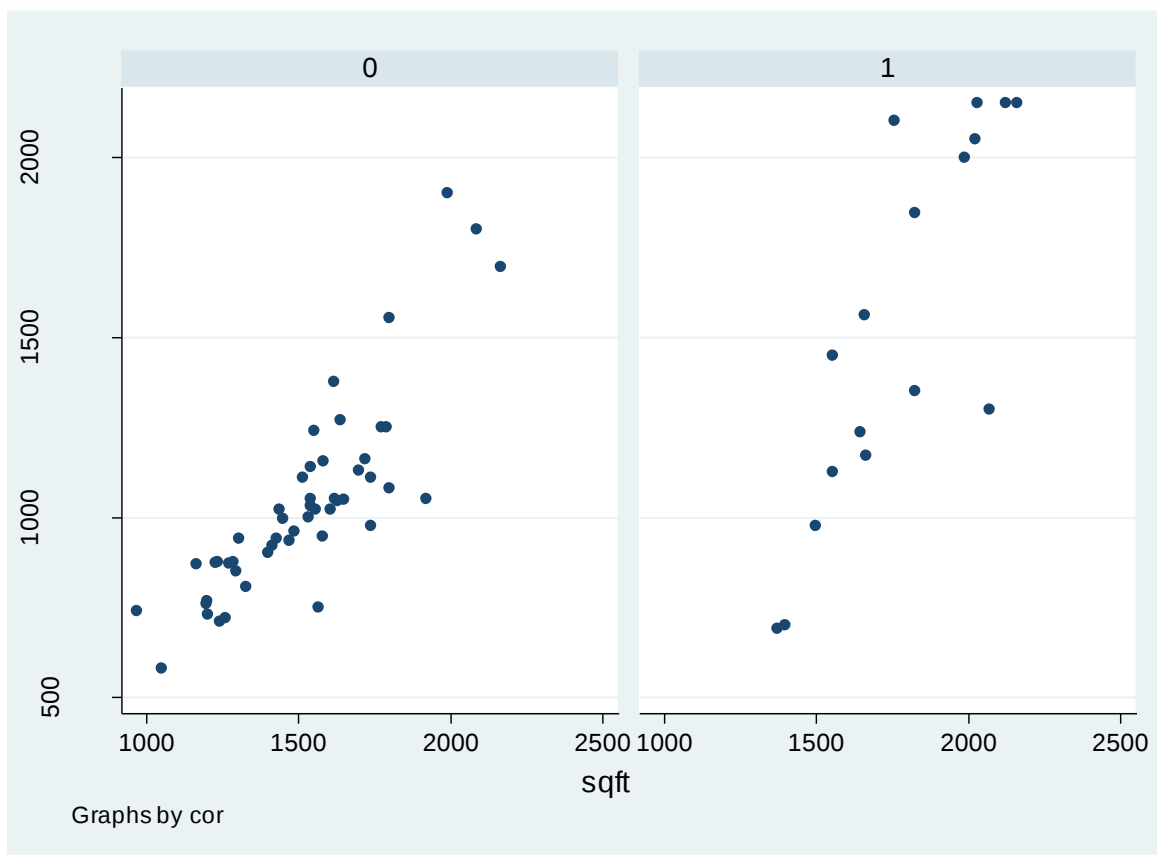
DATA Explore Bivariate – Stata commands

Descriptive statistics

sort cor

by cor: corr price sqft

scatter price sqft, by(cor)

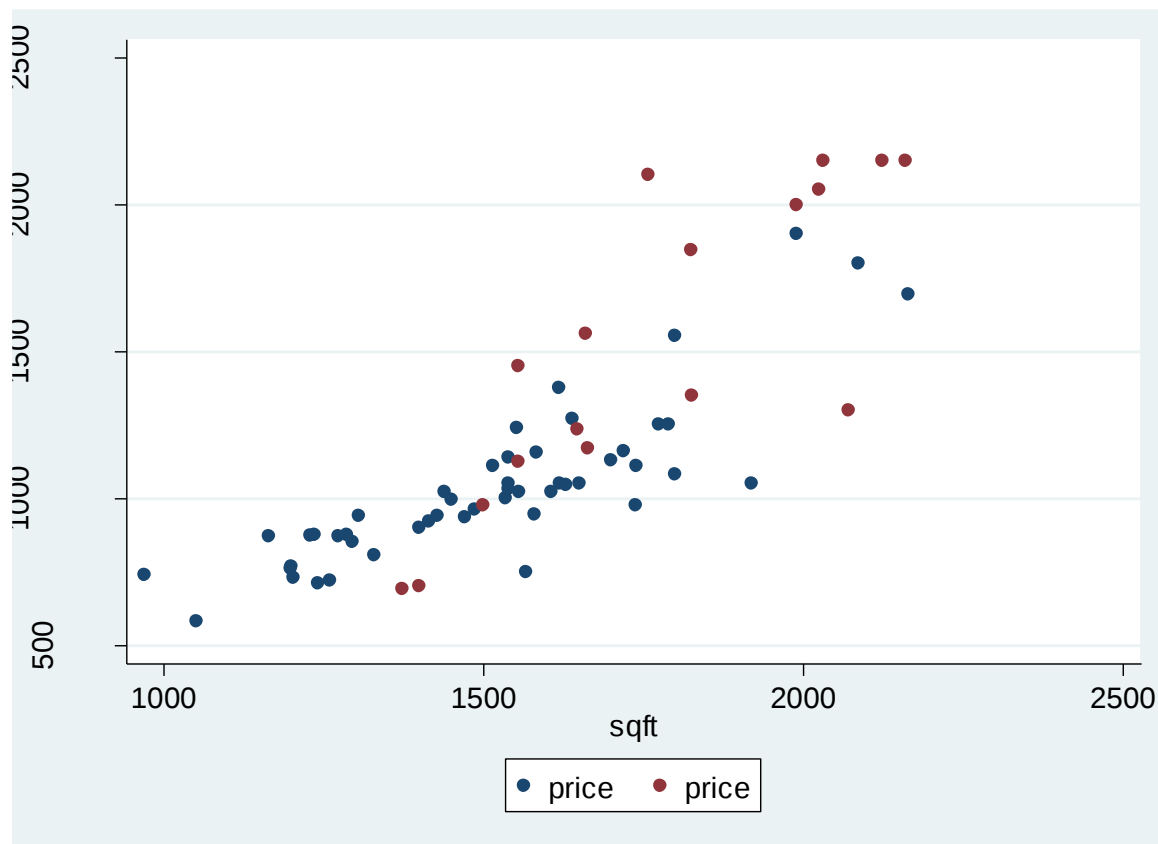


DATA Explore Bivariate – Stata commands

Descriptive statistics

twoway (scatter price sqft if cor==0) (scatter price sqft if cor==1)

by cor: reg price sqft



DATA Explore Bivariate – Stata commands

```
-----
-> cor = 0
Source |   SS   df    MS              Number of obs =   50
-----+-----              F( 1, 48) = 120.32
Model | 2523442.92   1 2523442.92      Prob > F   = 0.0000
Residual | 1006687.5   48 20972.6563      R-squared   = 0.7148
-----+-----      Adj R-squared = 0.7089
Total | 3530130.42   49 72043.478      Root MSE   = 144.82
-----
```

```
price |   Coef.   Std. Err.   t   P>|t|   [95% Conf. Interval]
-----+-----
sqft | .8840891   .0805983   10.97   0.000   .7220353   1.046143
_cons | -304.4202  123.9457   -2.46   0.018  -553.6296  -55.21076
-----
```

```
-----
-> cor = 1
Source |   SS   df    MS              Number of obs =   17
-----+-----              F( 1, 15) = 33.23
Model | 2924967.32   1 2924967.32      Prob > F   = 0.0000
Residual | 1320134.44   15 88008.9628      R-squared   = 0.6890
-----+-----      Adj R-squared = 0.6683
Total | 4245101.76   16 265318.86      Root MSE   = 296.66
-----
```

```
price |   Coef.   Std. Err.   t   P>|t|   [95% Conf. Interval]
-----+-----
sqft | 1.665617   .2889202    5.76   0.000   1.049798   2.281435
_cons | -1426.617   517.73   -2.76   0.015  -2530.132  -323.1014
-----
```



Example – Output

```
. reg price age feats ne cor sqft
```

Source	SS	df	MS	Number of obs =	67
Model	8478759.27	5	1695751.85	F(5, 61) =	43.54
Resid	2375649.38	61	38945.0718	Prob > F =	0.0000
Total	10854408.7	66	164460.737	R-squared =	0.7811
				Adj R-squared =	0.7632
				Root MSE =	197.35

	price	Coef.	Std. Err.	t	P> t	[95% Conf.Interval]
age	age of the building					
feats	number of options (reception, parking, garden,...)					
ne	if the house is in the north side (1) or not (0)					
cor	if the house is in the corner of building (1) or not (0)					
sqft	square feet of living space					
_cons	intercept					

Example – Output

```
. reg price age feats ne cor sqft
```

Source	SS	df	MS	Number of obs =	67
Model	8478759.27	5	1695751.85	F(5, 61) =	43.54
Resid	2375649.38	61	38945.0718	Prob > F =	0.0000
Total	10854408.7	66	164460.737	R-squared =	0.7811
				Adj R-squared =	0.7632
				Root MSE =	197.35

	price	Coef.	Std. Err.	t	P> t	[95% Conf.Interval]
age	-5.712149	2.121296	-2.69	0.009	-9.953942	-1.470356
feats	4.935951	21.39437	0.23	0.818	-37.84473	47.71663
ne	146.4015	60.79612	2.41	0.019	24.83218	267.9709
cor	191.3786	61.85902	3.09	0.003	67.68387	315.0734
sqft	.9871056	.1010516	9.77	0.000	.7850405	1.189171
_cons	-478.1104	162.0919	-2.95	0.005	-802.2331	-153.9876