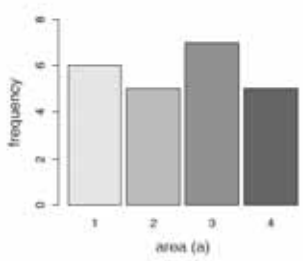
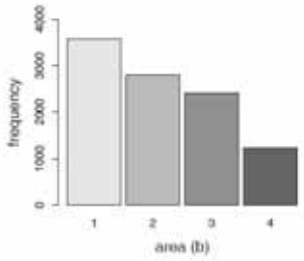
Machine Learning
Exploratory Data Analysis

1

- univariate analysis
  - Categorical
  - Numerical
- pairwise analysis
- multivariate analysis

Graphical analysis of categorical attributes





empirical frequencies

$$f_h = \frac{e_h}{m} \quad h \in \mathcal{H}$$

number of obs in h / total

for large samples

$$f_h \approx \Pr\{x = v_h\} \quad h \in \mathcal{H}$$

2

Analysis of categorical attributes

homogeniety / hetrogeniety (each class has same probability)

GINI INDEX

$$G = 1 - \sum_{h=1}^{H} f_h^2$$

ENTROPY INDEX

$$E = -\sum_{h=1}^{H} f_h \log_2 f_h$$

heterogeneity max

$(H-1)/H$ $\log_2 H$

0 0

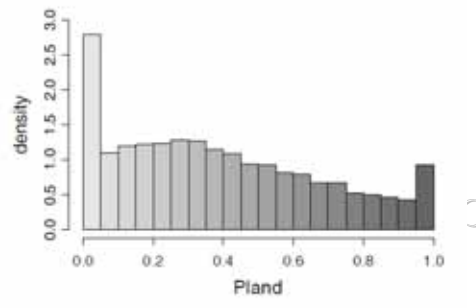
heterogeneity min

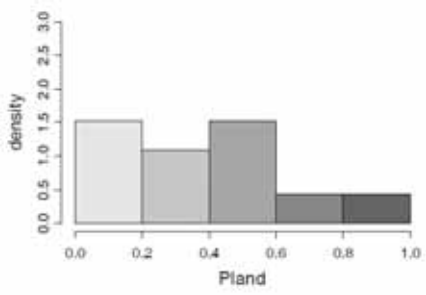
third in classification will be introduced

3

Analysis of numerical attributes

HYSTOGRAM





$$p_r = \frac{e_r}{ml_r}$$

4

Measures of central tendency for numerical attributes

هرکدوم بدی خوبی خودشو داره.

وقتی داده‌های خیلی بزرگ و کوچک داری خیلی اثر میذارن.

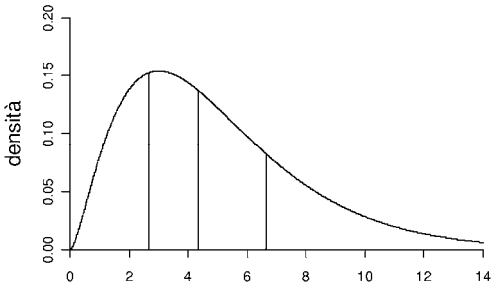
SAMPLE ARITHMETIC MEAN:

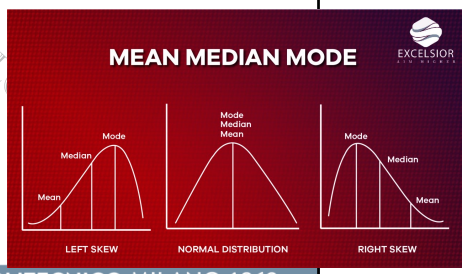
$$\bar{\mu} = \frac{x_1 + x_2 + \cdots + x_m}{m} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

مدیان معمولا جاش ثابته و روباسته

MEDIAN: central value of the observations

MODE: value that corresponds to the peak of the empirical density curve

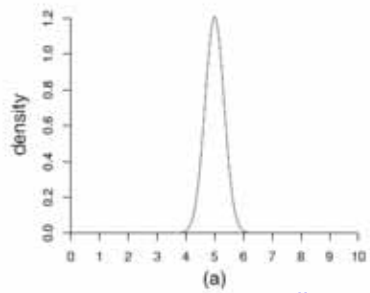


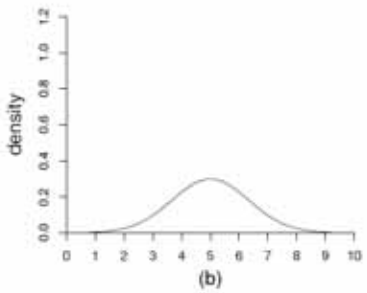


5

اما این ها کاملا نمایش دهنده وضعیت داده ها نیستند. در تصویر زیر هر سه مورد یکی خواهد بود.

Measures of dispersion for numerical attributes





errors are small.

6

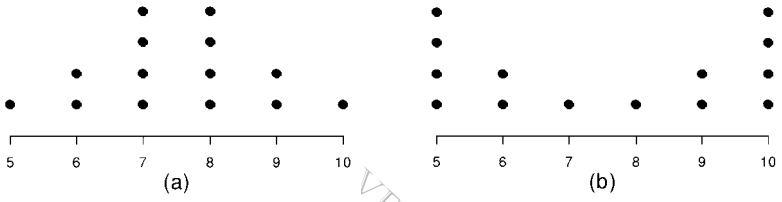
Measures of dispersion for numerical attributes

RANGE:

$$x^{range} = x^{max} - x^{min}$$

$$x^{max} = \max_i x_i$$

$$x^{min} = \min_i x_i$$



again it's not enought
both of them are same range.

7

Measures of dispersion for numerical attributes

other measures.

MEAN ABSOLUTE DEVIATION:

$$\text{MAD} = \frac{1}{m}\sum_{i=1}^{m}|x_i - \bar{\mu}|$$

SAMPLE STANDARD DEVIATION:

$$\bar{\sigma} = \sqrt{\bar{\sigma}^2}$$

نیاز داریم که ببینیم اسکیل داده ها چقدره؟

این بیشتر ترجیح داده شده

SAMPLE VARIANCE:

$$\bar{\sigma}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{\mu})^2$$

COEFFICIENT OF VARIATION:

$$\text{CV} = 100\frac{\bar{\sigma}}{\bar{\mu}}$$

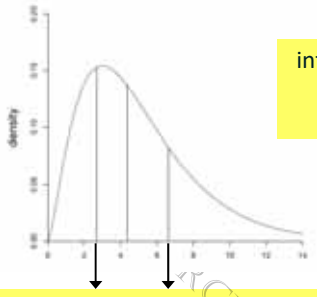
نشون میده که چقدر پخشه داده ها

8

Measures of relative location for numerical attributes

p ORDER QUANTILE q_p

leaev p percent left and the left to other side



interquartile distance

$D_q = q_U - q_L$ the smaller the better

$q_L = q_{0.25}$ $q_{0.75} = q_U$

9

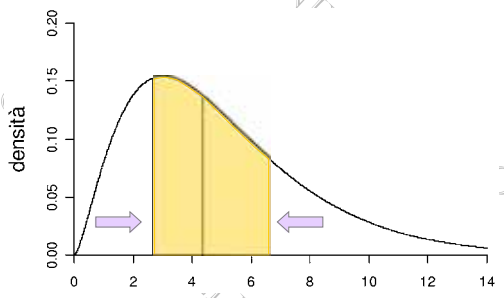
improvement over the sample mean, idea is to cut off in order to disregards the high affecting data

Measures of relative location for numerical attributes

MEAD-MEAN: mean of the values between q_L e q_U

TRIMMED-MEAN: mean of the values between q_P e q_{1-p}

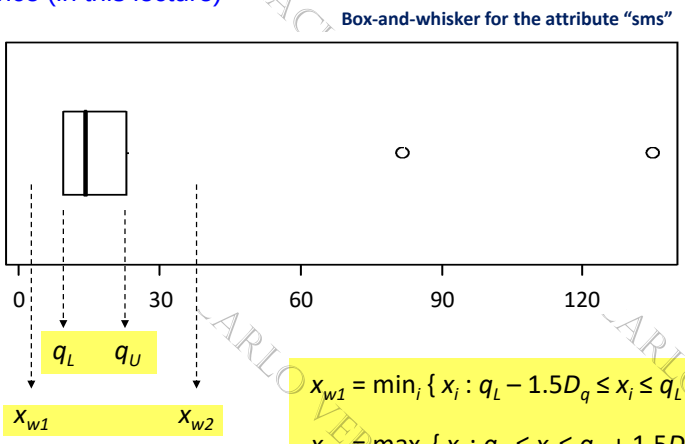
WINSORIZED-MEAN: increase (decrease) values less (greater) than the quantile q_P (q_{1-p})



10

Box-and-whisker

identifying outliers
assessing importance (in this lecture)



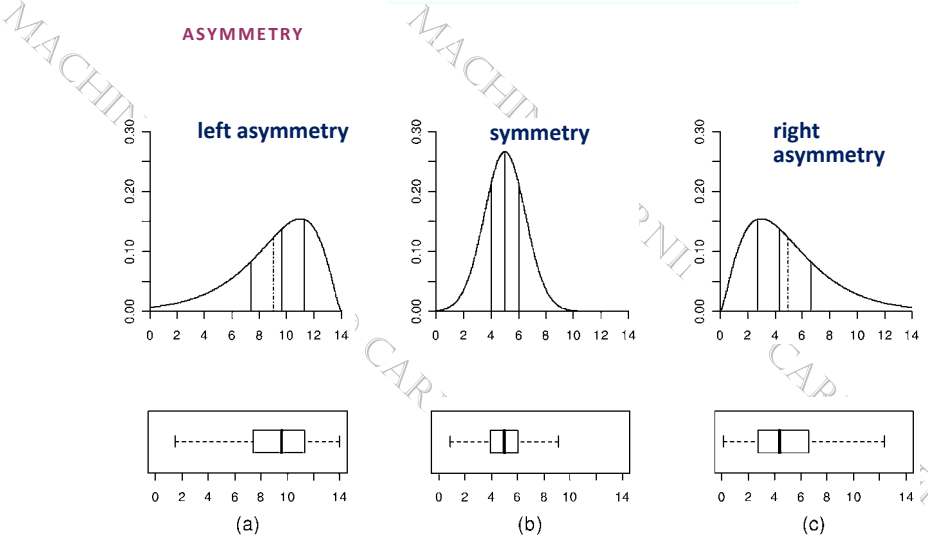
$$x_{w1} = \min_i \{ x_i : q_L - 1.5D_q \le x_i \le q_L \}$$

$$x_{w2} = \max_i \{ x_i : q_U \le x_i \le q_U + 1.5D_q \}$$

11

Analysis of the empirical density

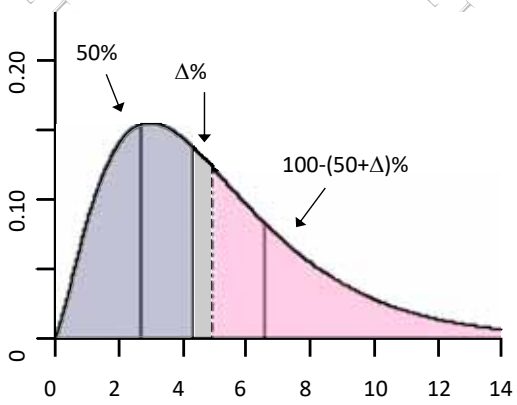
symmetry vs asymmetry



12

Analysis of the empirical density

ASYMMETRY **index :** $I_{as} = \frac{\bar{\mu}_3}{\bar{\sigma}^3}$ n $\quad \bar{\mu}_3 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{\mu})^3$



$I_{as} > 0$ **right asymmetry**

$I_{as} < 0$ **left asymmetry**

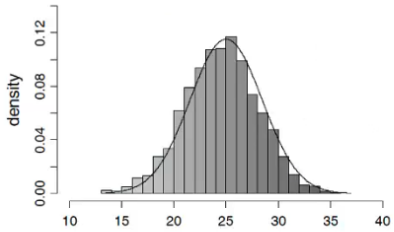
$I_{as} = 0$ **symmetric**

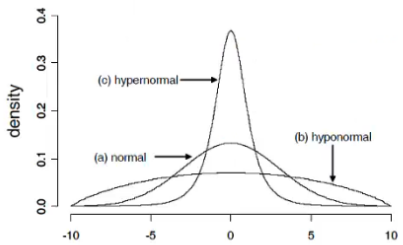
13

Analysis of the empirical density

KURTOSIS

$$I_{\text{curt}} = \frac{\bar{\mu}_4}{\bar{\sigma}^4} - 3 \qquad \bar{\mu}_4 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{\mu})^4.$$





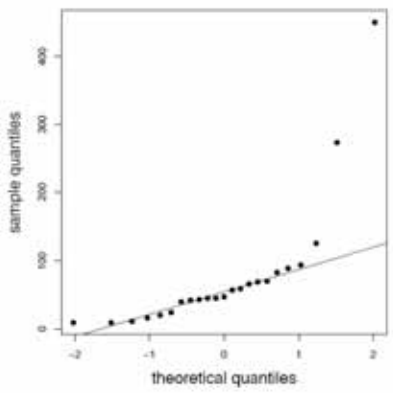
14

Analysis of the empirical density

tests for assessing normality

search how to interpret this.

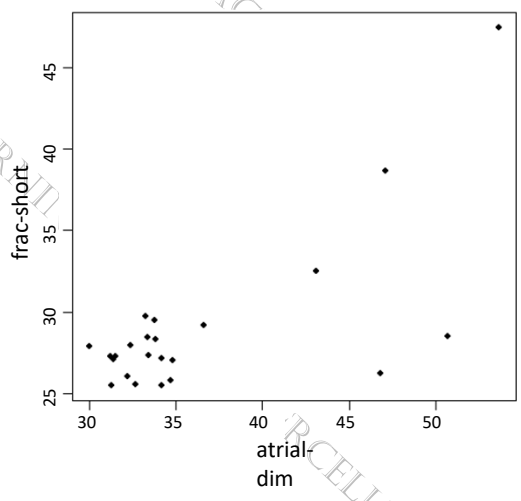
NORMAL-PROBABILITY PLOT



15

Bivariate Analysis --> can be either numerical or categorical

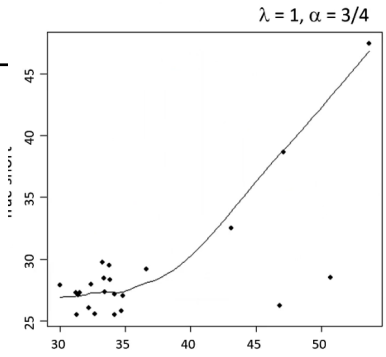
Scatterplot

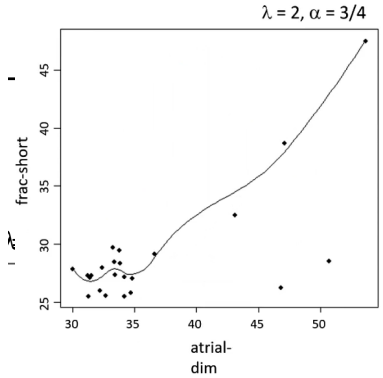


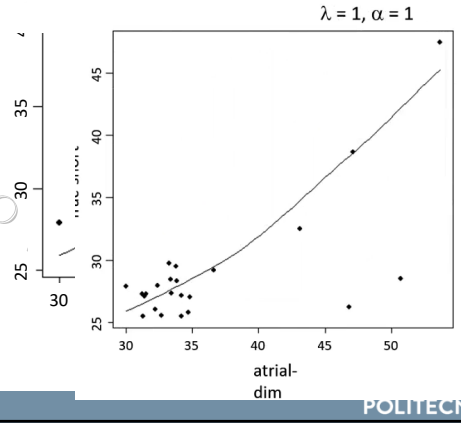
16

Loess diagram

loca regression technique







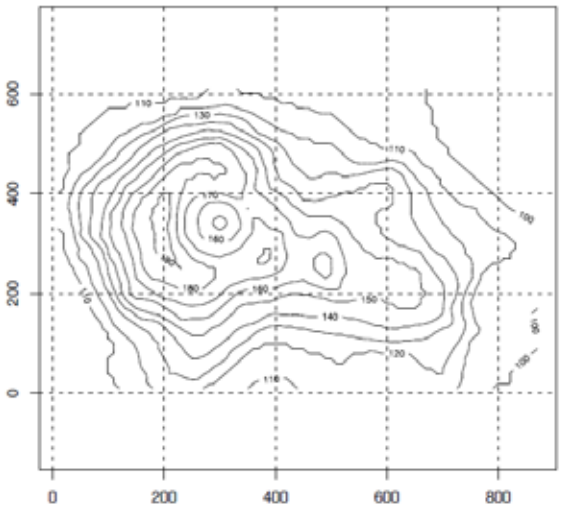
λ α

Regularization constant for neighborhood size

Degree of the polynomial for local regression

17

Contour lines

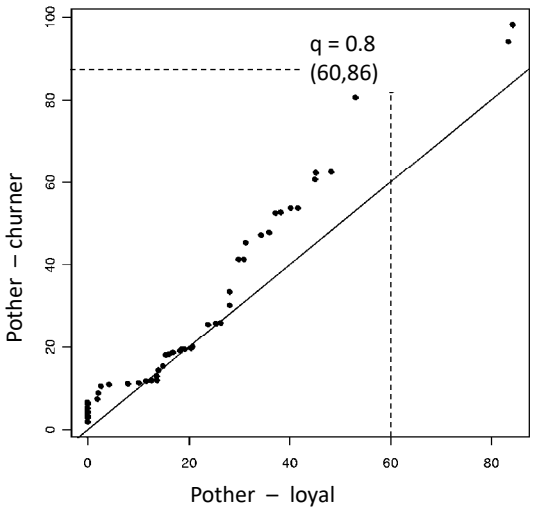


18

similar to normal probablity

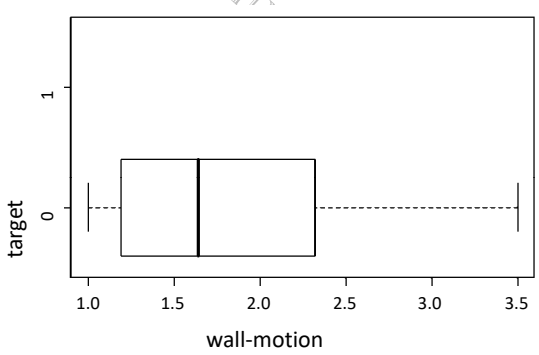
these two variable have same ~~distribution with another one? we~~ are taking the same variable!

Quantile-quantile plot



19

Box-and-whisker



20

statistical

Measures of correlation for numerical attributes

SAMPLE COVARIANCE:

$$v_{jk} = \text{cov}(\mathbf{a}_j, \mathbf{a}_k) = \frac{1}{m-2}\sum_{i=1}^{m}(x_{ij} - \bar{\mu}_j)(x_{ik} - \bar{\mu}_k)$$

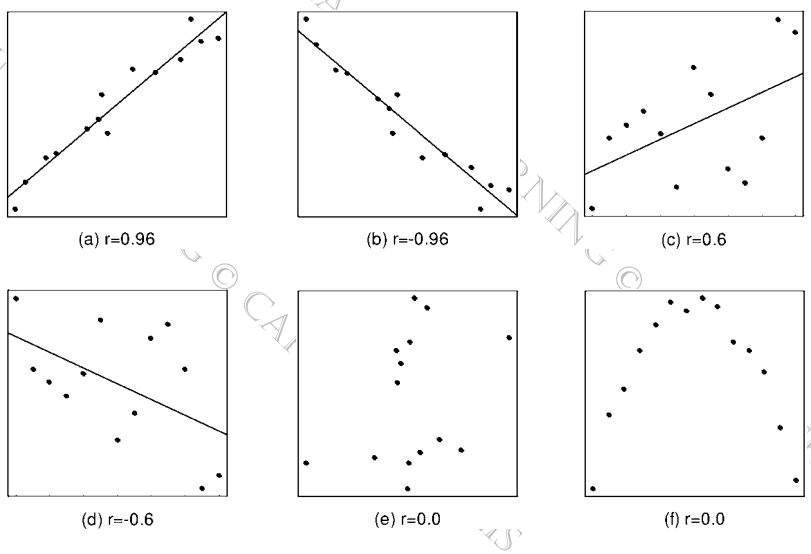
between two col

SAMPLE LINEAR CORRELATION :

$$r_{jk} = \text{corr}(\mathbf{a}_j, \mathbf{a}_k) = \frac{v_{jk}}{\bar{\sigma}_j \bar{\sigma}_k} \qquad r_{jk} \in [-1, 1]$$

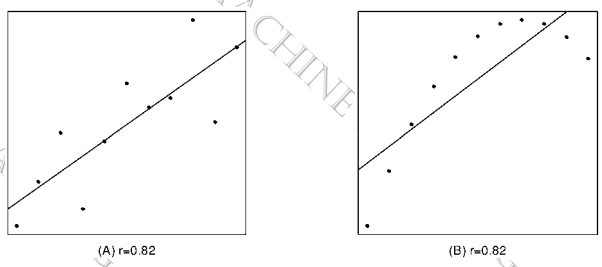
21

Sample linear correlation coefficient



22

Sample linear correlation coefficient



23

Contingency table

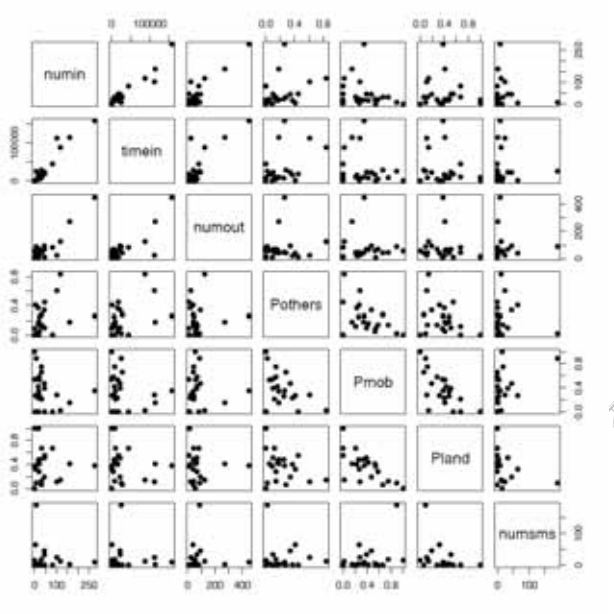
	family		
area	0	1	totale
1	2	4	6 (f_1)
2	4	2	6
3	2	5	7
4	3	3	6
totale	11 (g_1)	14 (g_2)	25

Two attributes are independent if

$$\frac{t_{r1}}{g_1} = \frac{t_{r2}}{g_2} \quad r = 1, 2, \ldots, J$$

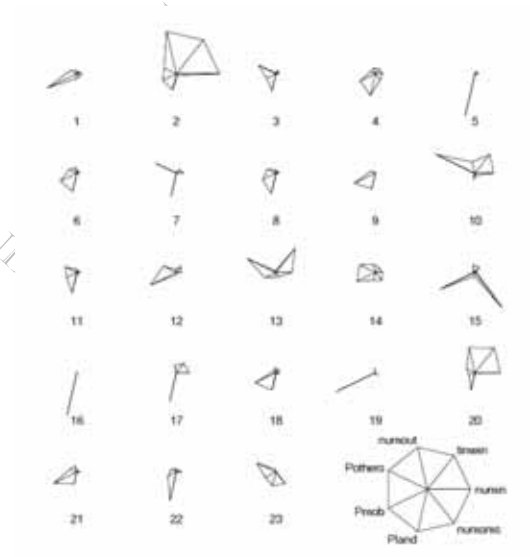
24

Scatter plot matrix



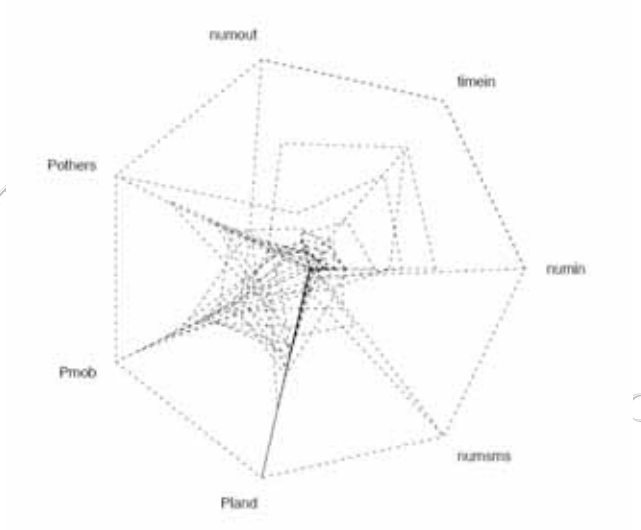
25

Star plots



26

Spider web chart



27

Measures of correlation for numerical attributes

covariance	numin	timein	numout	Pothers	Pmob	Pland	numsms
numin	3769.99	2433848.00	5689.04	2.35	−2.50	0.07	−233.25
timein	2433848.00	1997570000.00	3522037.00	1893.14	−2518.33	575.88	−44328.60
numout	5689.04	3522037.00	12525.06	−2.90	0.10	2.76	92.83
Pothers	2.35	1893.14	−2.90	0.04	−0.02	−0.02	0.63
Pmob	−2.50	−2518.33	0.10	−0.02	0.06	−0.04	0.37
Pland	0.07	575.88	2.76	−0.02	−0.04	0.06	−1.00
numsms	−233.25	−44328.60	92.83	0.63	0.37	−1.00	657.63

correlation	numin	timein	numout	Pothers	Pmob	Pland	numsms
numin	1.00	0.89	0.83	0.19	−0.17	0.00	−0.15
timein	0.89	1.00	0.70	0.21	−0.24	0.05	−0.04
numout	0.83	0.70	1.00	−0.13	0.00	0.10	0.03
Pothers	0.19	0.21	−0.13	1.00	−0.41	−0.42	0.12
Pmob	−0.17	−0.24	0.00	−0.41	1.00	−0.65	0.06
Pland	0.00	0.05	0.10	−0.42	−0.65	1.00	−0.16
numsms	−0.15	−0.04	0.03	0.12	0.06	−0.16	1.00

28