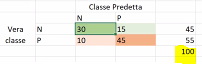
*Type 1 Error* → falso positivo

*Type 2 Error* → falso negativo, ad esempio un soggetto che è positivo dico che è negativo

In statistica c’è la matrice di confusione che consente di mettere in relazione positivi e negativi.

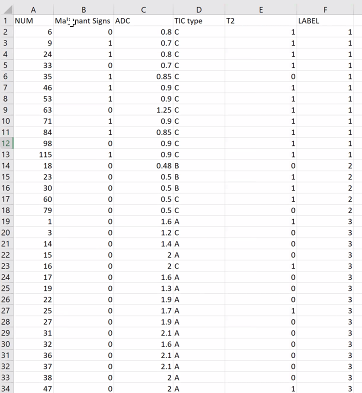
Ci sono i classificatori che consentono di predire la classe di appartenenza di un soggetto (i valori vengono riportati nella matrice di confusione).

La somma dei valori in tabella deve dare 100.

Sulla riga **N** si hanno i veri **n**egativi e sulla riga **P** si hanno i veri **p**ositivi.

I falsi positivi sono 15 → si è detto che sono positivi ma in realtà sono negativi; mentre i falsi negativi sono 10.

Dataset di esempio

Quando si vede un dataset, la prima cosa da fare è capire le variabili presenti.

• ***Malign Signs*** è una variabile booleana.

• **ADC** è una variabile continua.

• ***TIC type*** è una variabile categorica non ordinale che assume 4 valori.

Le variabile categoriche possono anche essere ordinabili, come i mesi dell’anno.

Esiste un ordinamento implicito ossia pensare che febbraio sia maggiore di gennaio.

Anche l’altezza di un individuo è una variabile categorica ordinabile.

• **T2** è una variabile booleana.

• Si hanno poi ***Label*** e ***Num***, che identifica il singolo paziente.

Come si fa er visualizzare queste variabili?

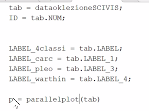
Per visualizzare ***Malignant Signs*** che è una variabile booleana si può utlizzare un istogramma.

Da una **label** che ha 4 classi si è creata una label binaria “**Label\_1**”, che rende più semplice la visualizzazione.

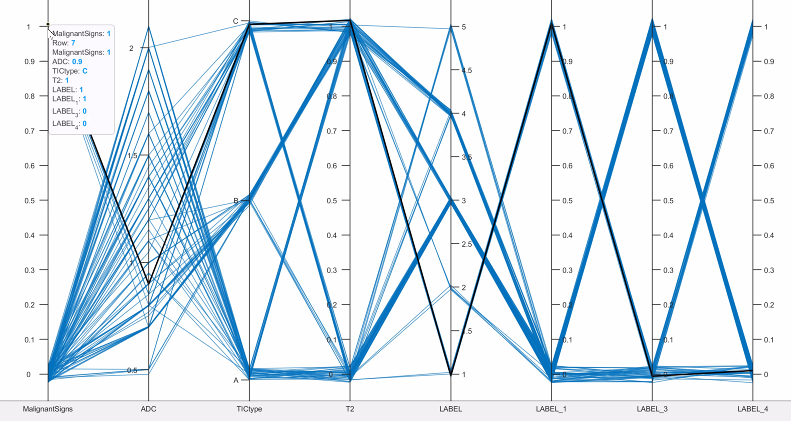
Si può visualizzare ogni classe rispetto alle altre, quindi si creano **Label\_3** e **Label\_4** che fanno riferimento alle variabili 3 e 4.

(**MATLAB** ha la stessa logica di **Matplotlib** e viceversa.)

In MATLAB:



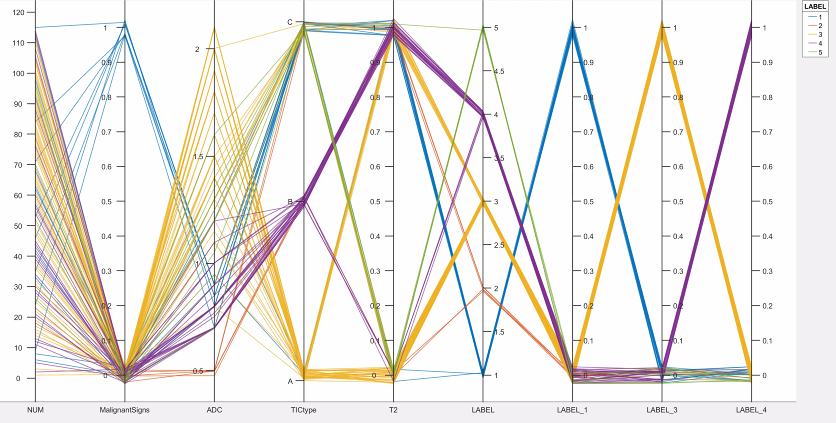
Per ogni punto si crea una linea che scorre tutte le variabili nel dataset:



Ogni linea del **parallel plot** è un paziente.

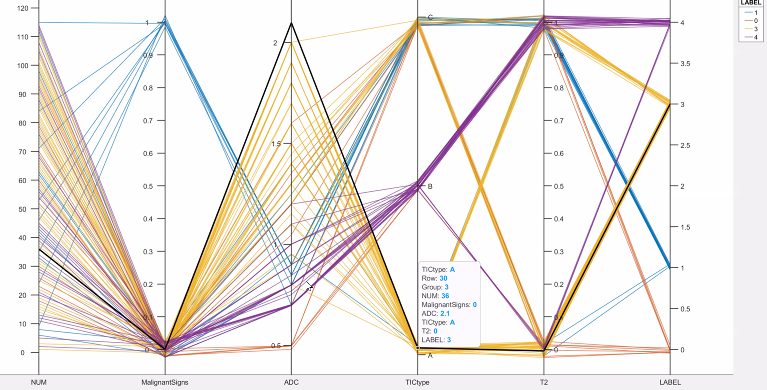
Per raggruppare: 

Le linee sono state raggruppate in base alle classi trovate in LABEL:

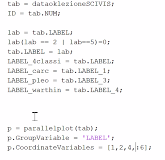


Il colore è associato alla classificazione del tumore del paziente: 

Si stanno eliminando due classi, la 2 e la 5, dato che non ci interessano e quindi sono state messe a 0.

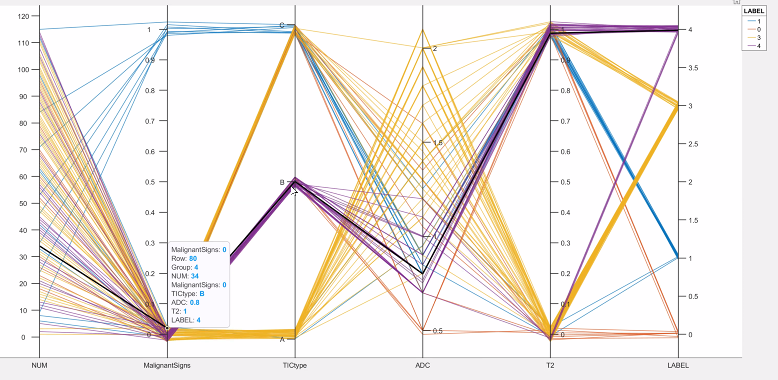


Per mettere in risalto la classe viola:





**figure()** per stampare il plot (l’equivalente di **plt.figure** in Matplotlib).



Il ***parallel plot*** è utile quando si hanno variabili categoriche (per vedere come si distribuiscono) e si vogliono poche variabili continue.

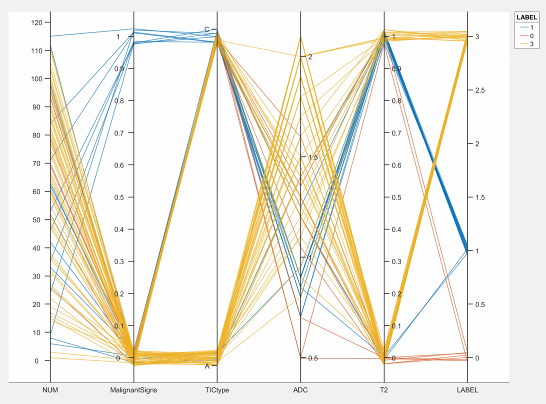
Si creano degli assi verticali, ognuno associato ad una variabile diversa, e ogni linea che rappresenta il paziente congiunge i diversi valori.

Quando si capisce come è distribuita una data classe (avendola osservata dal *parallel plot* e avendo compreso nella totalità il plot), quella classe può anche essere rimossa dal dataset (avendo compreso che magari non è necessaria).

(La tabella si può fare allo stesso modo in ***Pandas***.)

Si rimuove la classe viola (dato che si è visto non essere rilevante):





La visualizzazione interattiva permette di conoscere meglio il dataset.

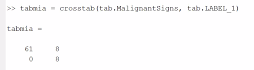
Come si visualizza però ogni singola variabile contro la label?

Si deve considerare il tipo della label.

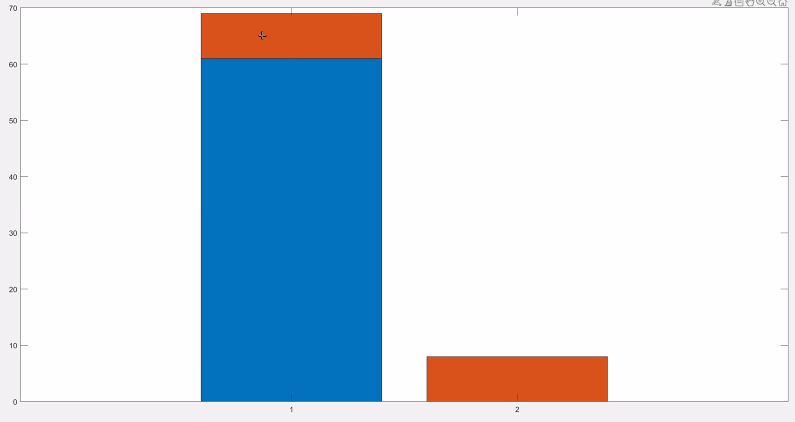
***Malignant Signs***, ***TIC type*** e **T2** sono categoriche quindi si utilizza un istogramma semplice a due bin, considera le tre classi (per ogni classe si può fare un istogramma per ***Malignant Signs***, ***TIC type*** e **T2**).

Un istogramma associa una label ad ogni barra e se si considera ad esempio ***Malignant Signs*** si vedrà nella barra quanti sono i pazienti che hanno valore 0 o valore 1 per ***Malignant Signs***.

Si vede la variabile ***Malignant Signs*** contro la label:



Lo ***stacked bar chart*** fa vedere per due diversi gruppi della label com’è la distribuzione all’interno della classe (***Malignant Signs***) che si sta visualizzando: 

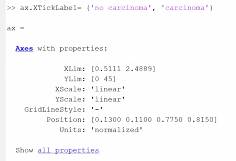


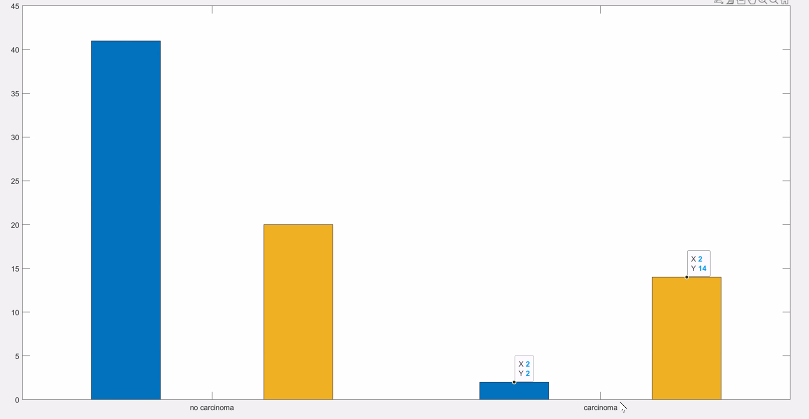
Le variabili categoriche si possono vedere come sono distribuite rispetto ad una classe.

Ora si vede la variabile ***TIC type*** contro la label:



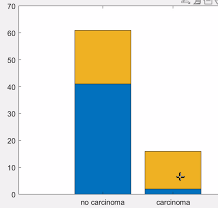


Sono state aggiunte le proprietà: 



Se si vogliono visualizzare delle variabili categoriche, si plottano le distribuzioni tramite dei ***bar chart***.

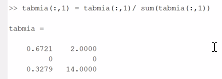
Versione “***stacked***” del bar chart:

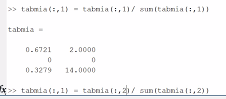
Si riesce a comparare bene le parti sotto (i blu si vedono bene che ce ne sono di più in “no carcinoma”) ma la classe gialla è più difficile confrontarla.

Sarebbe stato meglio normalizzare le due colonne rispetto a tutti i pazienti “no carcinoma”.

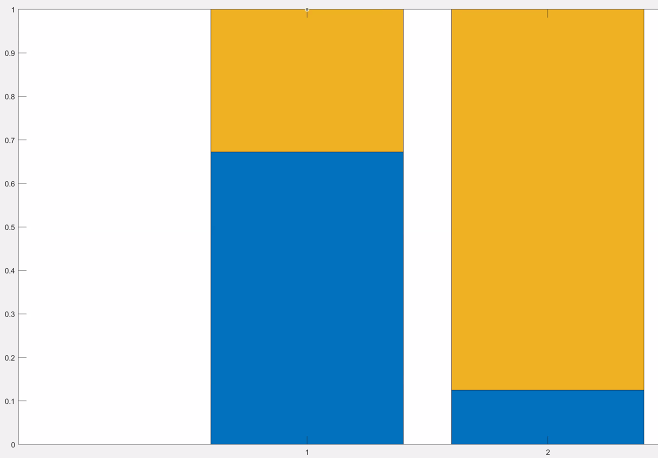
Quindi si poteva plottare la tabella, dopo averla normalizzata.

Si normalizza perciò la prima colonna della tabella rispetto alla somma della prima colonna (coloro senza carcinoma):

Si normalizza poi la seconda colonna:

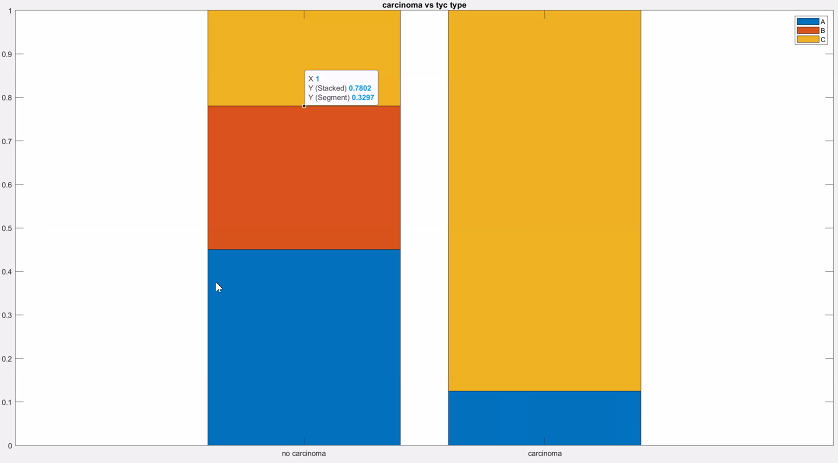


Risultato:



Da questo plot si può vedere come le persone senza carcinoma abbiano una curva **TIC** di tipo A.

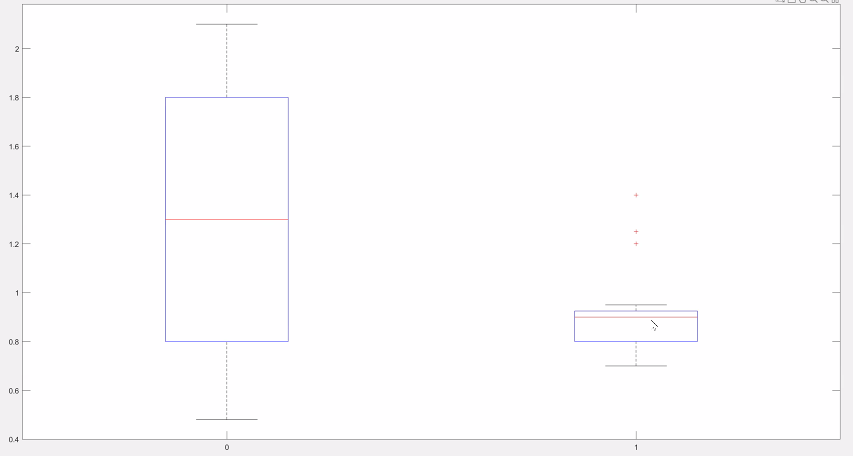




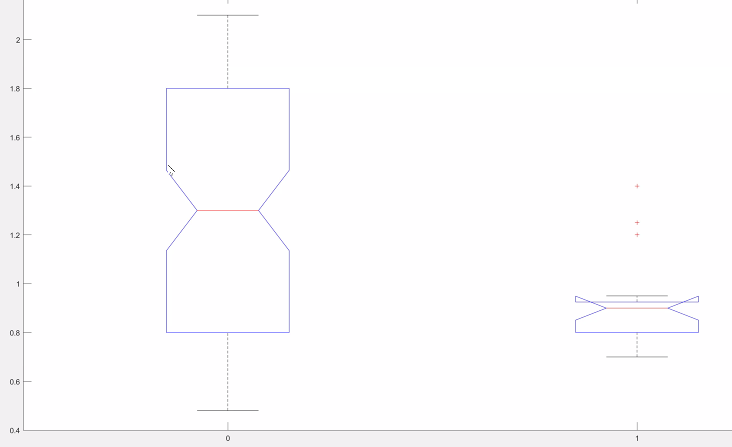
Da ↑ si evince che per coloro che hanno il carcinoma, non ci sono pazienti che hanno il TIC type uguale a B.

Per visualizzare invece la variabile continua **ADC** si crea un ***box plot***: il box plot stampa la distribuzione.

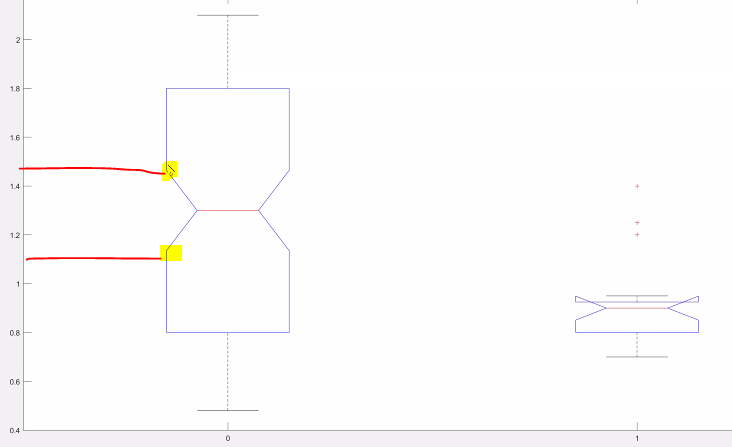
La linea rossa indica la media dei valori dell’**ADC** che hanno valori pari a 0.



Si possono fare diversi tipi di ***box plot***, tipicamente si fa quello con il ***notch***:



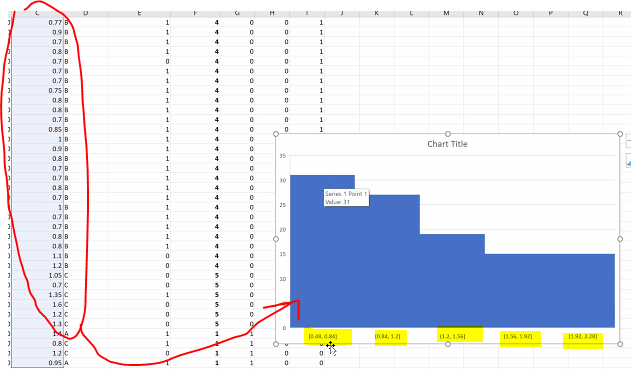
L’intervallo di confidenza della media o della mediana è 1.18 più o meno 1.41:

La mediana ha un intervallo di confidenza più stretto rispetto alla media.

Si hanno degli *outlier*, che sono i punti rossi, ossia coloro che non hanno dei valori di **ADC** comuni agli altri (quei pazienti si differenziano di gran lunga rispetto agli altri per quanto riguarda il valore di **ADC**) .

Per le varibaili categoriche si utilizzano i ***bar chart*** e se si vuole capire l’andamento delle variabili categoriche si utilizzano i ***parallel plot***.

Slide 32 (30 blocco slide Lezione6.pdf)



Excel ha creato 5 bin, ha diviso i valori della colonna evidenziata in azzurro in gruppetti, in modo da poterli porre sul ***bar chart***.

***Stacked Density Plot***

L’idea di ***GapMinder*** è stata quella di mettere vicino colori simili.

Tutti i rossi nel ***density plot*** **sovrapposto** sono sovrapposti ma posti vicini.

Questa disposizione dei colori dipende dal principio della Gestalt.

Slide 33 (31)

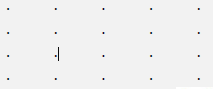
I principi della Gestalt sono tanti, noi vediamo quelli relativi alle visualizzazioni.

Principio di prossimità: elementi che sono visualizzati vicini nel campo di visione tendono ad essere percepiti come appartenenti allo stesso gruppo, mentre elementi lontani nel campo di visione tendono ad essere percepiti come appartenenti a gruppi differenti.

Da questo principio nascono i ***grouped bar chart***.



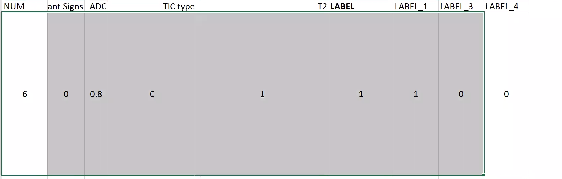
Si fa in modo che le colonne della tabella siano piccole e le righe siano alte, così da indurre il lettore a fare un confronto tra le righe.

Se iveece si vuole indurre il lettore a far confrontare le colonne, si distanziano di più le colonne.

Se si uniscono le colonne non si capisce molto, se però si distanziano tanto, quello che viene allora da fare è un confronto in verticale:



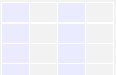
Successivamente viene poi fatto il confronto in orizzontale, ponendo la tabella in questo modo:

si restringono le colonne e si alzano le righe.

Slide 34 (32)

Principio di similarità: oggetti che sono simili per colore, dimensione o forma sono percettivamente visti come raggruppati.

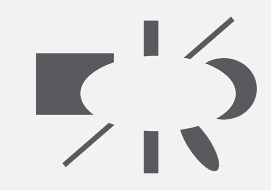
Nelle tabella si possono avere colori alternati per righe e colonne.

Si può utilizzare uno stesso colore per diverse colonne, per indicare che queste appartengono ad una stessa catgeoria.

Si usano delle forme per imporre delle categorie: 

**N.B.** E’ ottimale avere colori nettamente separati tra di loro e non troppe forme (altirmenti è difficile ricordare le varie categorie).

Slide 35 (33)

Principio di chiusura: oggetti che sono chiusi in qualche modo tendono ad essere percepiti come gruppi.

Qui si hanno:

* il principio della chiusura perché tutti i punti all’interno dell’ellisse vengono visti come facenti parte di un gruppo
* il principio della continuità.

Slide 36 (34)

È stato utilizzato il principio della similarità per i colori, per andare a marcare determinare regioni.

Il colore indica la via di sviluppo della regione.

Se non fosse stata messa la linea tratteggiata non si avrebbe colto il rapporto alto PIL-bassa mortalità.

**Rosling** ha creato la chiusura (i cerchi tratteggiati) per far vedere il concetto generale: la chiusura identifica un’area che consente di raggruppare punti che hanno colore diverso.

Slide 35

La chiusura è anche il motivo per cui i grafici a barre, che creano aree chiuse in base al colore, sono quelli che secondo studi percettivi vincono su tutto. La chiusura dà una migliore percezione della dimensione.

I grafici a barre sfruttano area e dimensioni.

Principio di chiusura: tendiamo a chiudere forme continue.

Slide 36

Principio di continuità: ci fa percpepire due forme, ossia la linea ed il rettangolo, ma in realtà si ha un linea interrotta a tratti.

Questi **principi di Gestalt** possono essere utilizzati per le tabelle (vedi tabella in slide):

il principio di prossimità fa sì che si considerino sottotabelle, inoltre impone un confrono prima verticale (dato che le colonne sono assai distanziate) e poi orizzontale.

La continuità/prossimità fa si che non abbiamo bisogno dei bordi, in quanto vediamo bene le colonne.

Bar chart a DX: il principio di continuità/chiusura/similarità permette di dire che se si hanno della variabili categoriche, possiamo anche non mettere l’asse x, in quanto questo viene inferito dal punto 0 delle barre.

Dato che sull’asse x non si stanno plottando delle variabili continue, non è necessario metterlo.

Si potrebbe mettere il numero sopra la barra oppure utilizzare l’asse y.

Slide 37

Il fatto che ci sia una connessione tra le due coppie di punti fa percepire che quelle coppie di punti appartengano a gruppi differenti.

La connessione tende ad essere più forte della prossimità.

Da studi percettivi però la chiusura è la più forte.

La connessione è utile quando si vuol far percepire un trend e far vedere dei pattern.

Slide 38

È partito da anni precedenti ed ha plottato dei punti.

Slide 41

Cosa percepiamo?

Possiamo raggruppare in base all’*orientazione* ma non è precisa, in base alla *lunghezza* ma è difficile da deifnire.

Slide 43

Se si vuole identificare questi segmenti come dei punti, dove lo si mette il punto?

Al centro o alla fine.

Slide 44

Dato che non si riesce a raggruppare così i segmenti (posti come in slide 41), si possono utilizzare dei colori oppure il principio della chiusura.

Slide 46

Se si sostituisce tutto con una connessione, si connettono punti, li si fanno appartenere ad un gruppo e si fa vedere il trend.

Tutto ciò che si plotta ha un significato percettivo per noi.

Slide 49

Sono fondamentali anche i colori utilizzati nel testo.

Attenzione ad usare i colori nei testi, anche un *serif* può essere reso difficile da leggere.

Studi percettivi dicono che il *serif* più facile da leggere è il **Times New Roman**.

Slide 50

Le scritte si differenziano per la semplicità del carattere, ***Serif*** (con grazie – ghirigori), ***sans-serif*** (senza grazie), spaziatura tra i singoli caratteri…

Per diversi utenti la presenza di *serif* fornisce una più facile visualizzazione del testo.

Slide 52

In verde e rosso si vedono i diversi prezzi che salgono e scendono.

Il colore dipende dal periodo selezionata dall’utente.

Nelle visualizzazioni sfortunatamente non c’è una regola, dobbiamo vedere noi come disporre gli elementi e fare in modo che qualunque lettore capisca la nostra viusalizzazione.

Slide 54

Questa visualizzazione è pessima in quanto il lettore deve scorrere tutte le nazioni e non si capisce bene.

Se si sostituisce questa visualizzazione con un ***bar chart*** ↓

Slide 55

Nei bar chart si sfrutta il colore: verde indica tranquilità, rosso il negativo e grigio un qualcosa che si può ignorare.

Si possono plottare le barre a caso oppure in ordine alfabetico di ordinamento.

Se si hanno gruppi che non seguono un ordinamento logico (come le nazioni) si cerca comunque di dare un ordinamento.

Ad esempio nel bar chart a SX si potrebbero ordinare le nazioni in base al “favore” di queste nazioni rispetto agli USA.

In questo modo si è creato comunque un ordinamento.

L’azzurro indica che ci vuole una certa attenzione, ma che non si schiera da nessuna parte.