



Esercitazione OLAP

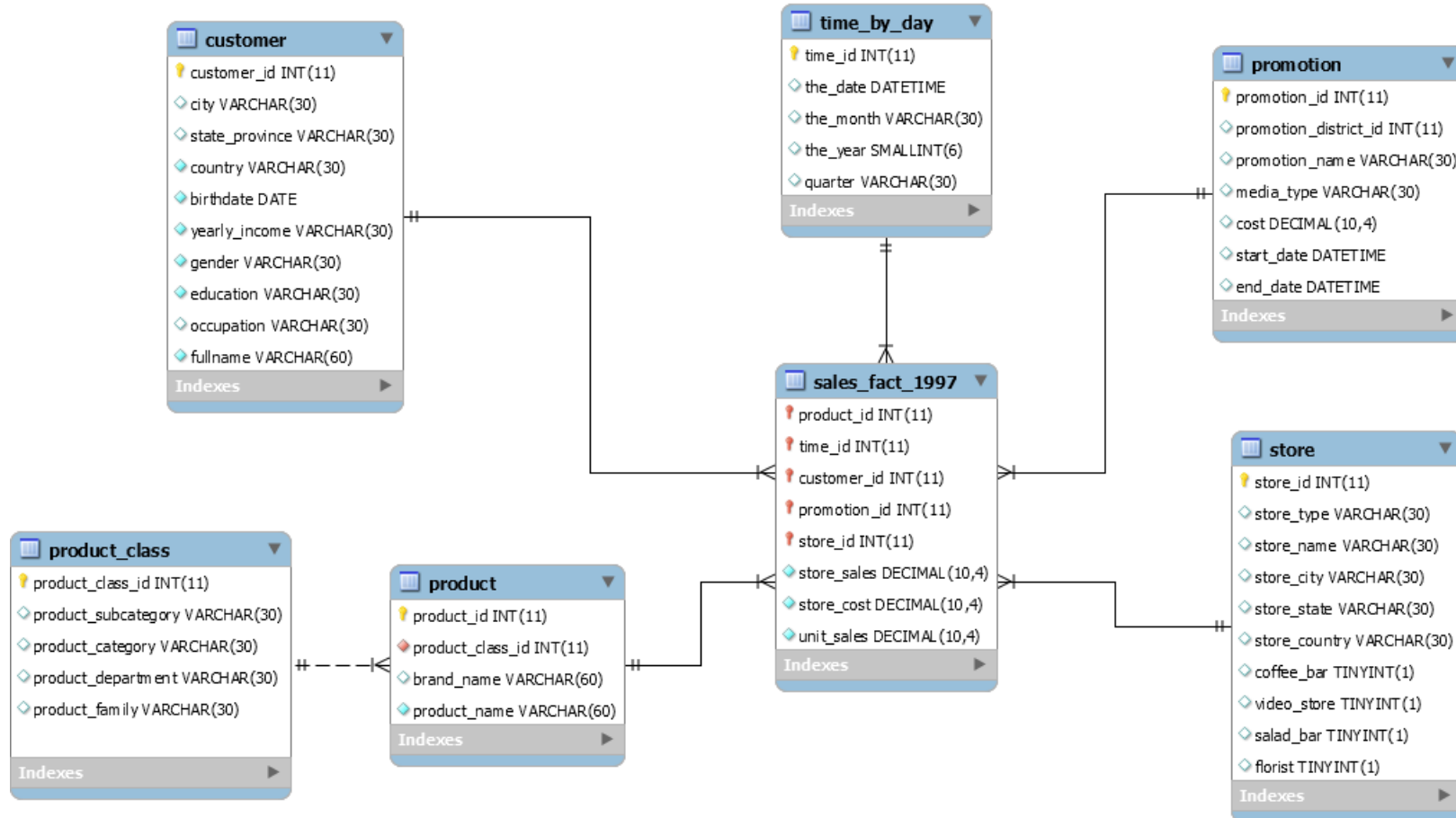
Master FCA, A.A. 2020/21

Matteo Francia

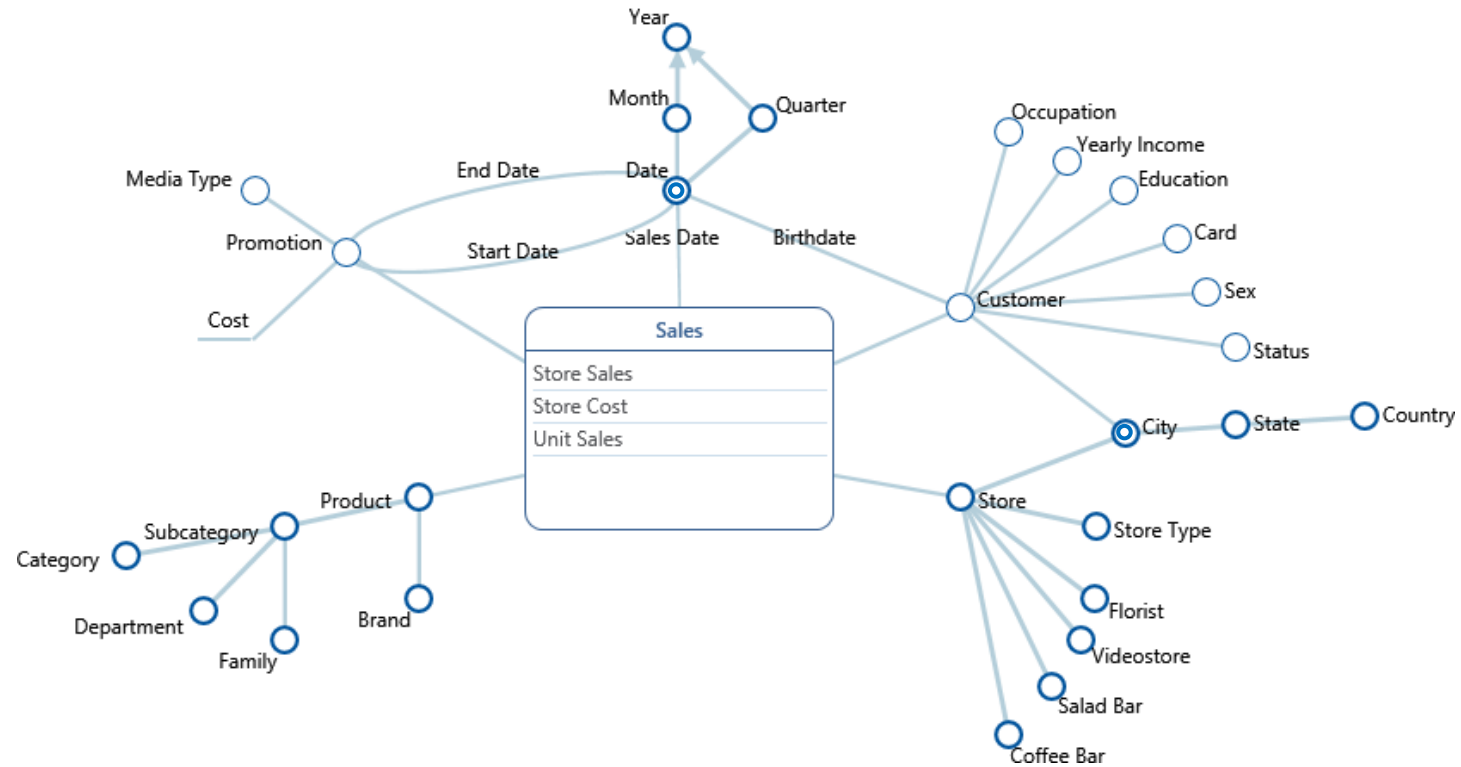
BOLOGNA BUSINESS SCHOOL
Alma Mater Studiorum Università di Bologna

UN BREVE RIPASSO

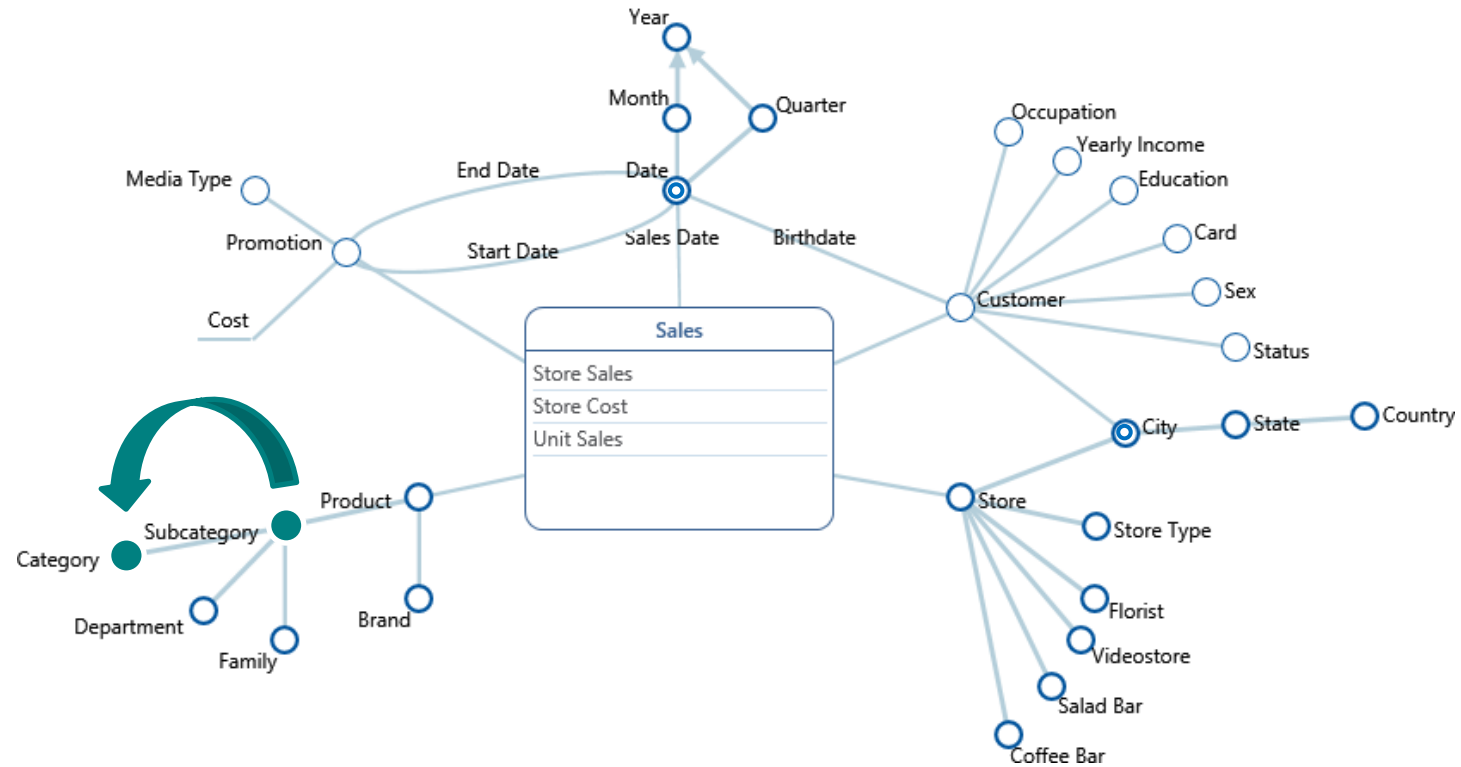
Sales (Foodmart) – Database Structure



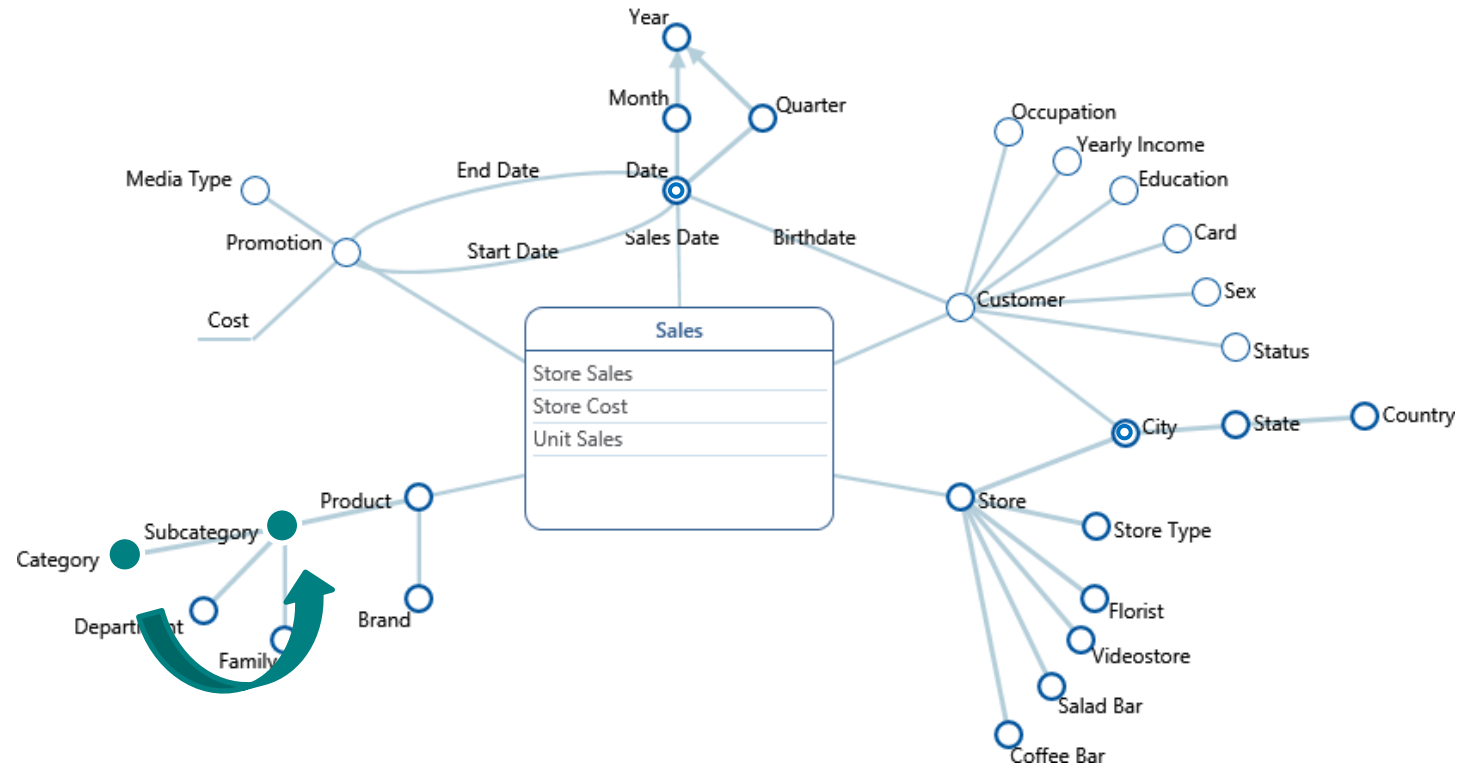
Sales (Foodmart) – DFM



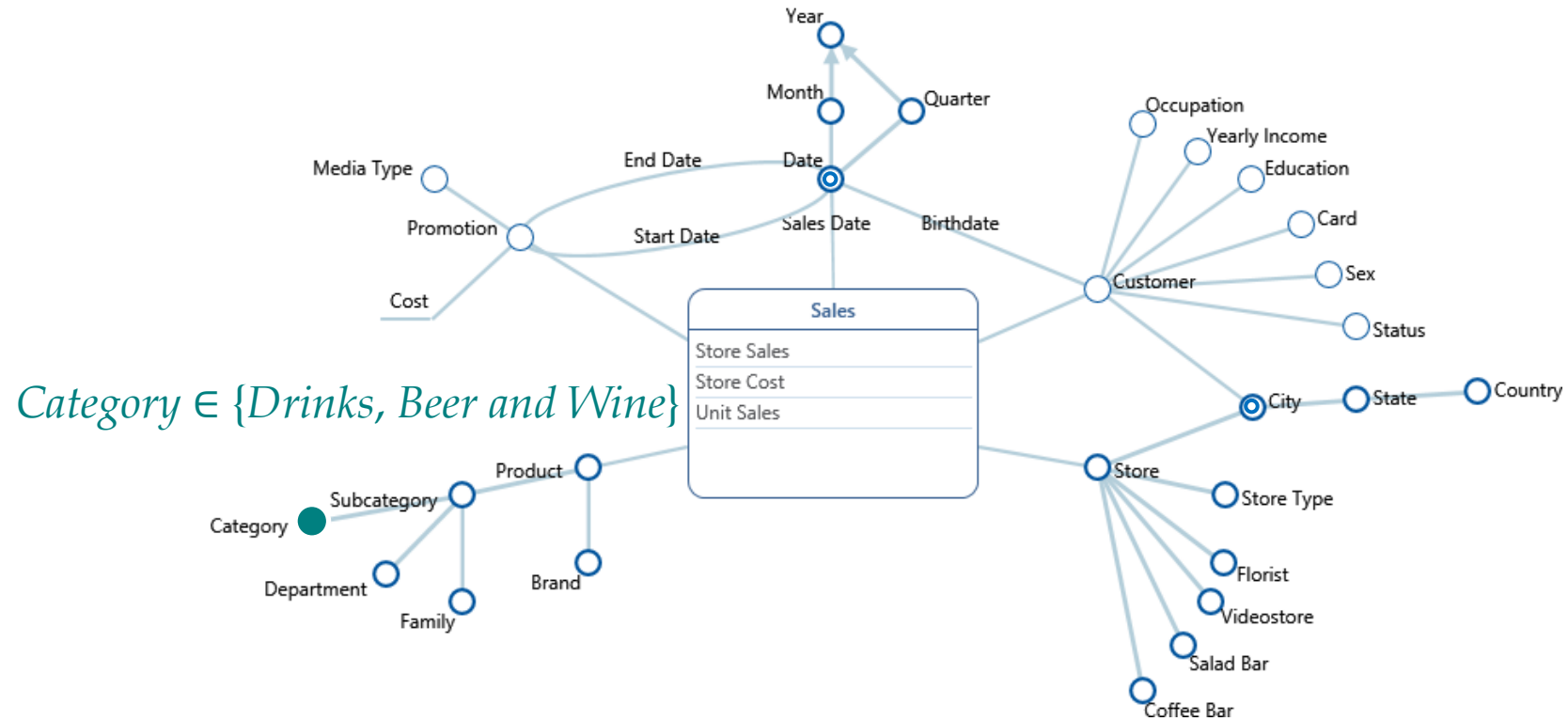
Operatori OLAP: Roll-Up



Operatori OLAP: Drill-Down

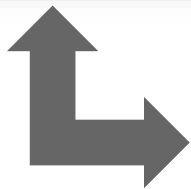


Operatori OLAP: Slice & Dice



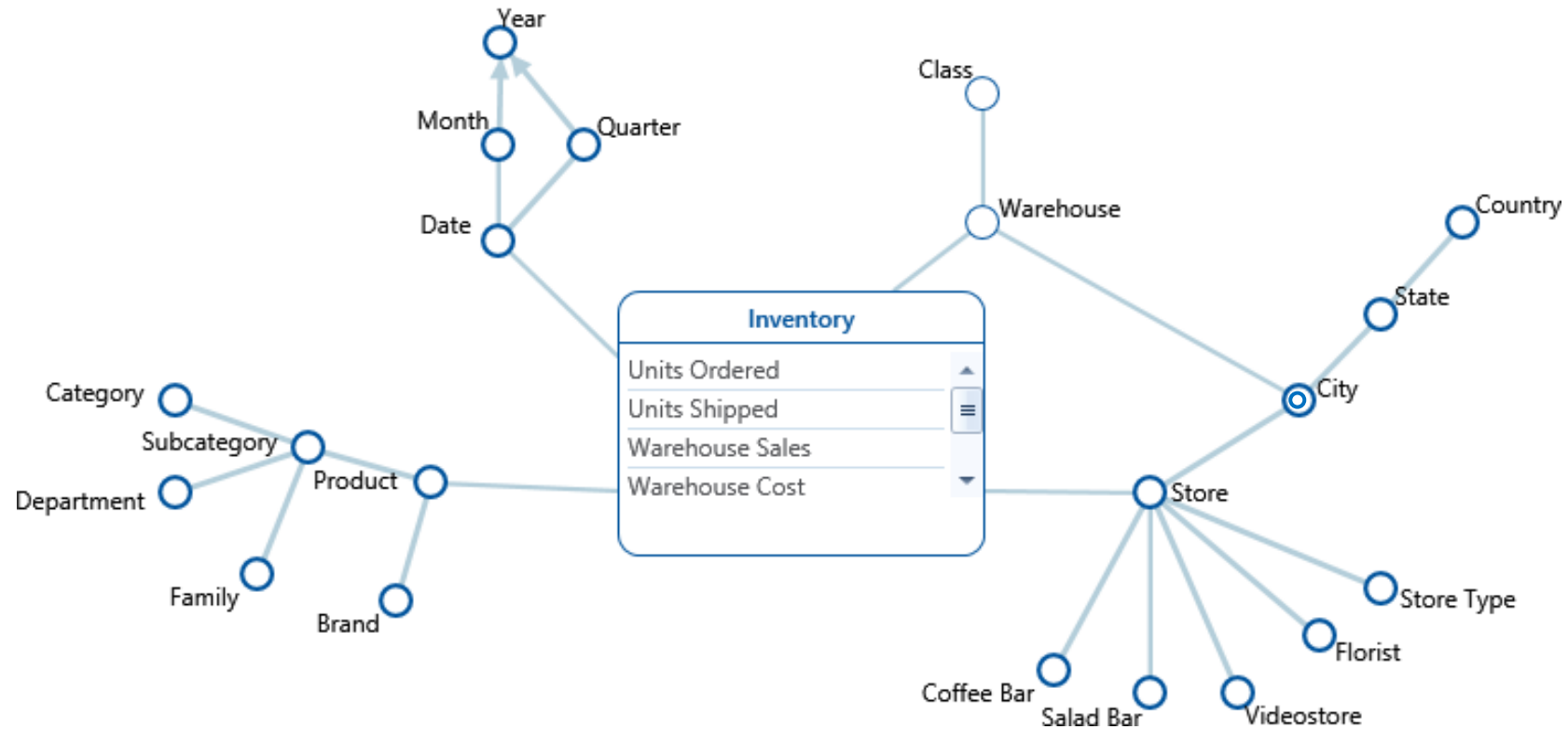
Operatori OLAP: Pivoting

Category	Occupation				
	Clerical	Management	Manual	Professional	Skilled Manual
Baking Goods	338	2,242	3,823	5,262	3,781
Bathroom Products	257	1,910	3,319	4,278	3,266
Beer and Wine	148	1,861	3,449	4,739	3,833
Bread	278	2,116	3,980	5,569	4,512
Breakfast Foods	239	2,405	3,855	4,879	4,179
Candles	10	231	304	463	353
Candy	188	2,117	3,699	4,636	3,910
Canned Anchovies	25	381	425	875	590
Canned Clams	6	229	445	774	459
Canned Oysters	19	236	427	495	265
Canned Sardines	7	172	404	436	338
Canned Shrimp	31	302	654	671	489
Canned Soup	251	2,207	3,908	5,531	4,068
Canned Tuna	23	444	751	1,114	879
Carbonated Beverages	126	974	1,333	2,107	1,696
Cleaning Supplies	97	986	1,715	2,365	1,951
Cold Remedies	59	454	817	1,275	752
Decongestants	28	464	806	1,172	830
Drinks	79	887	1,413	1,779	1,485



	Category													
Occupation	Baking ..	Bathroo..	Beer an..	Bread	Breakfa..	Candles	Candy	Canned ..	Canned ..	Canned ..	Canned ..	Canned ..	Canned ..	Canned ..
Clerical	338	257	148	278	239	10	188	25	6	19	7	31	251	23
Management	2,242	1,910	1,861	2,116	2,405	231	2,117	381	229	236	172	302	2,207	444
Manual	3,823	3,319	3,449	3,980	3,855	304	3,699	425	445	427	404	654	3,908	751
Professional	5,262	4,278	4,739	5,569	4,879	463	4,636	875	774	495	436	671	5,531	1,114
Skilled Manual	3,781	3,266	3,833	4,512	4,179	353	3,910	590	459	265	338	489	4,068	879

Inventory (Foodmart) – DFM



Inventory (Foodmart) – Database Structure

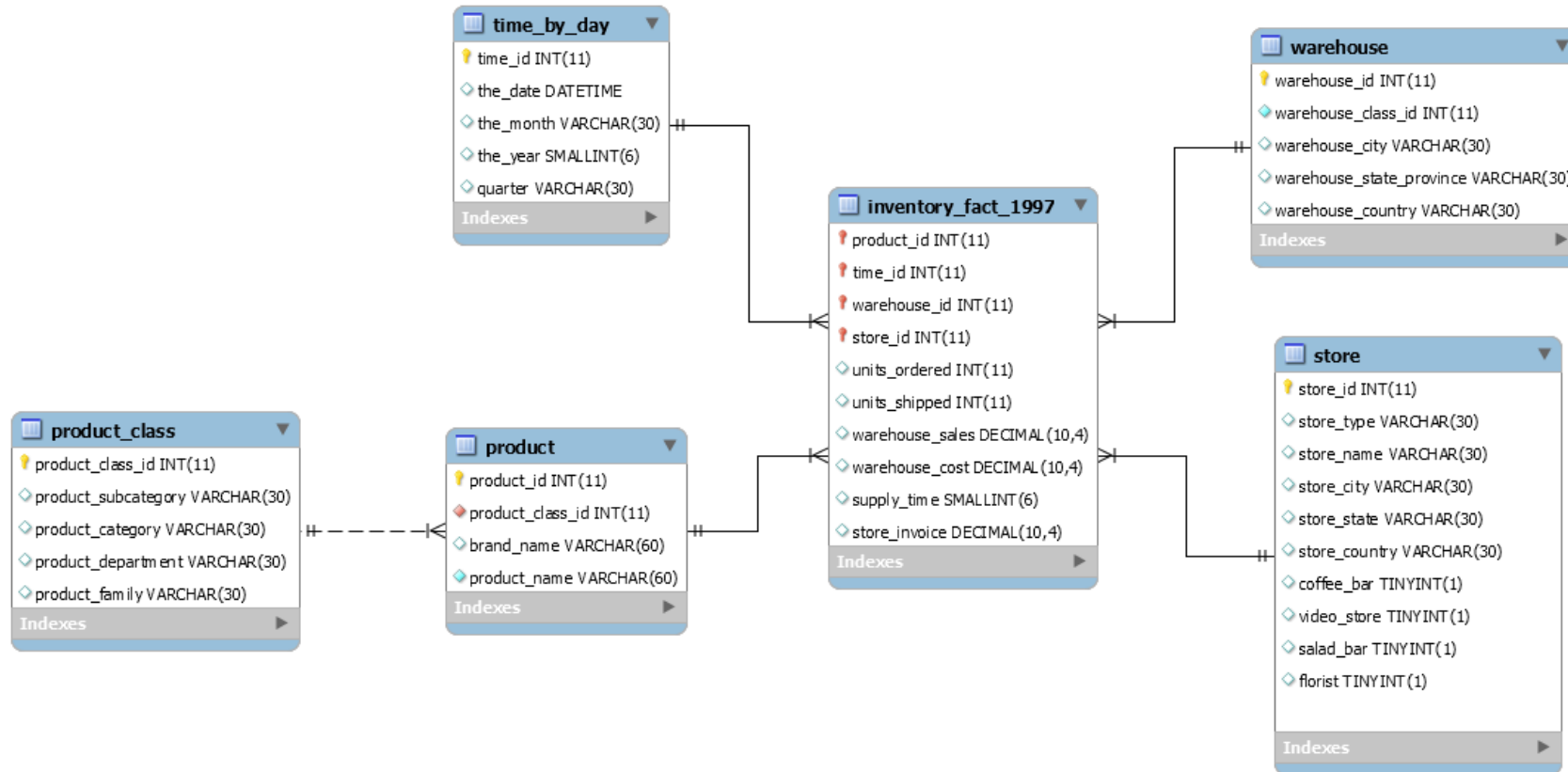
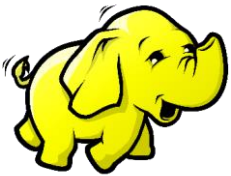
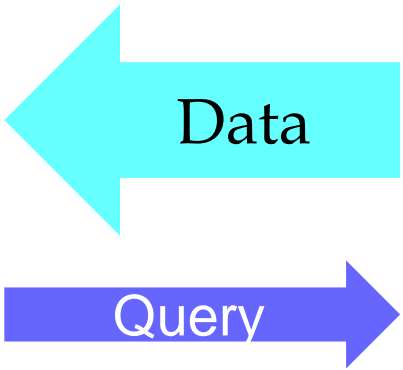


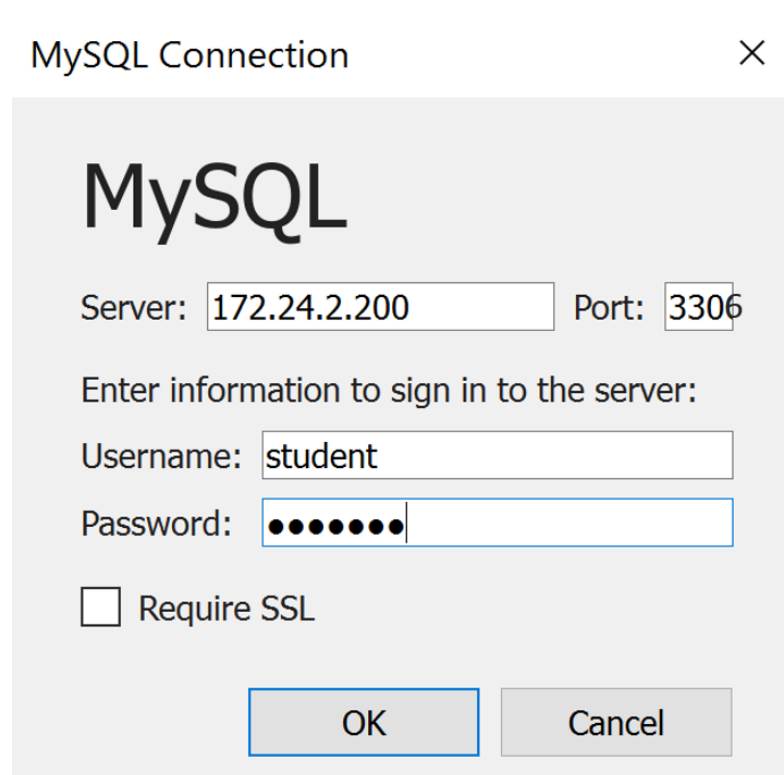
TABLEAU: VISUAL ANALYTICS

Architettura



Connessione

- | Scaricare il file **foodmart_sales.twbx** e aprirlo
- | Cliccare su **Data Source** in basso a sinistra e impostare la connessione come in figura
- | Aprire infine un foglio di lavoro (**Worksheet**)



MySQL Connection

MySQL

Server: 172.24.2.200 Port: 3306

Enter information to sign in to the server:

Username: student

Password: ●●●●●●

☐ Require SSL

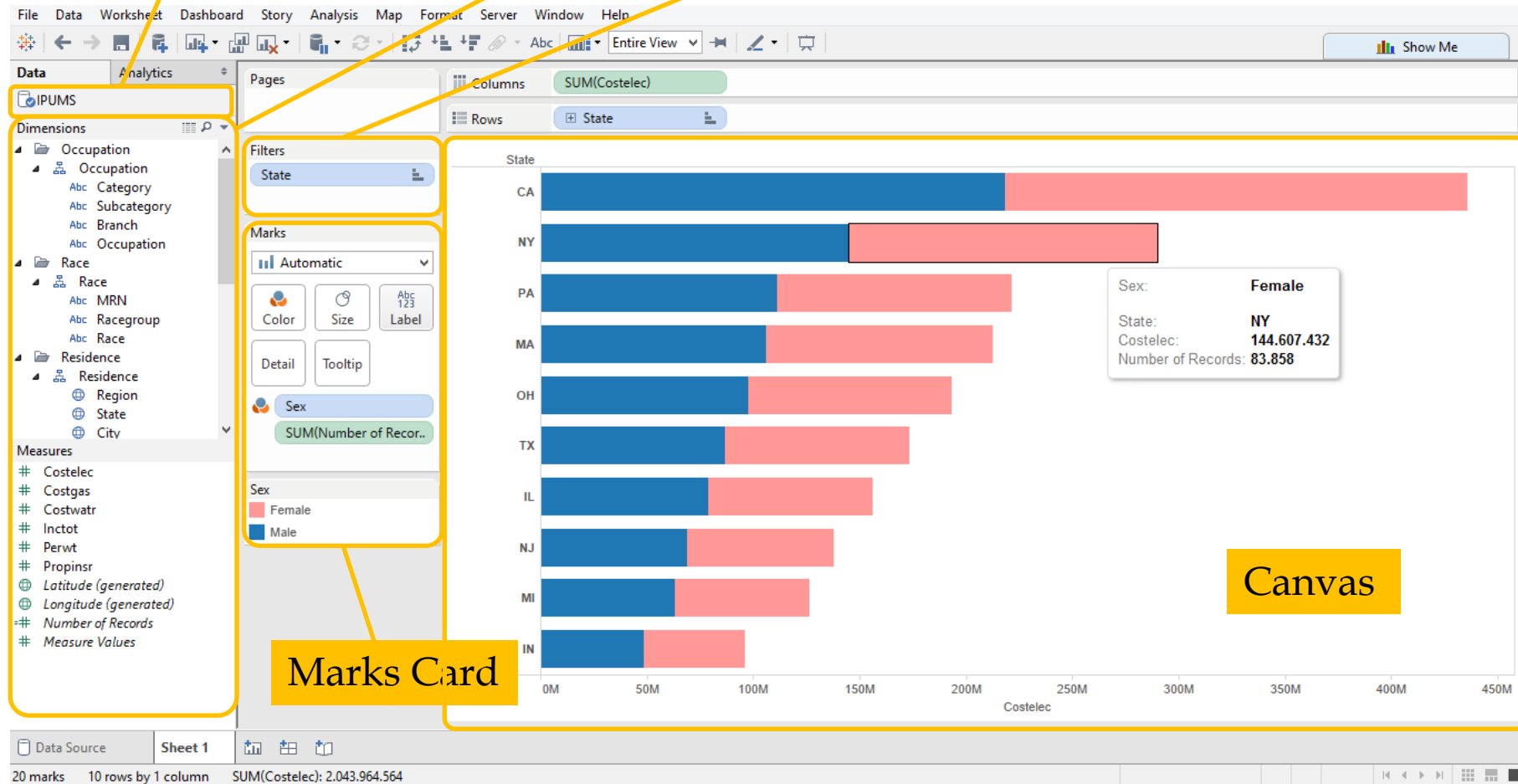
OK Cancel

Interfaccia

Sorgenti dati

Dimensioni e misure

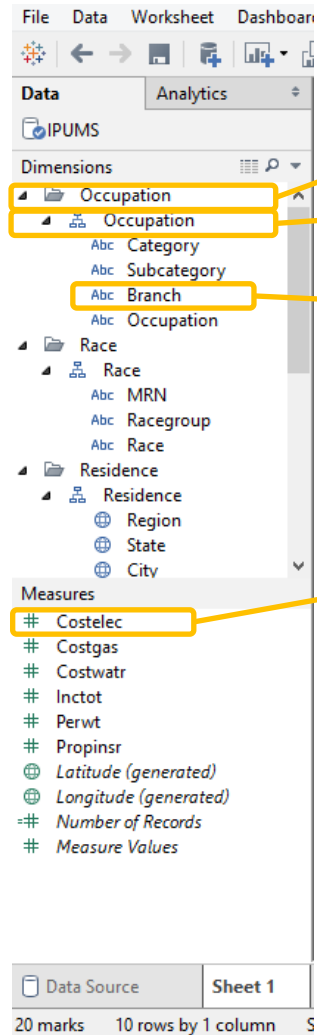
Filtri



Canvas

Marks Card

Interfaccia



Dimensione

Gerarchia

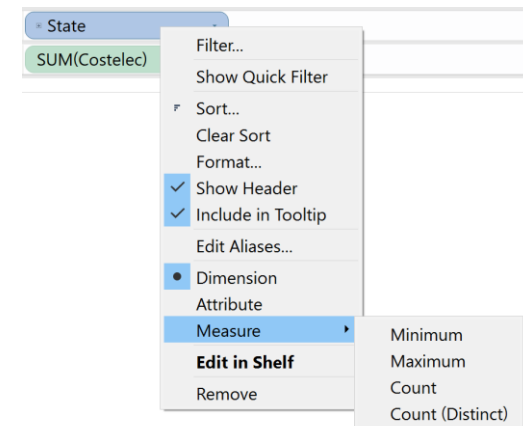
Attributo dimensionale

Misura

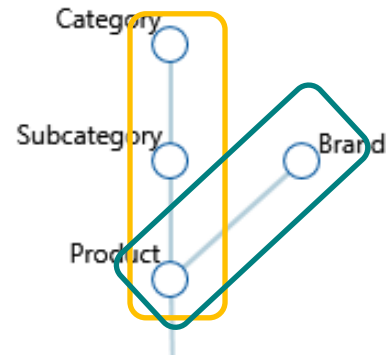
- Per convenzione, le **cartelle** vengono usate per rappresentare **dimensioni**; in generale è solo un modo per raggruppare elementi

Dimensione VS Misura

- | In Tableau le definizioni di *dimensione* e *misura* sono più lasche rispetto a quelle tradizionalmente utilizzate in letteratura e ogni campo può essere utilizzato sia come dimensione che come misura
- | In generale è comunque utile dare una classificazione iniziale ai campi
 - * Una dimensione è un qualunque campo **indipendente** (eg. *città*)
 - * Una misura è un qualunque campo i cui valori sono **funzione di altri campi** (eg. *profitto vendite*)



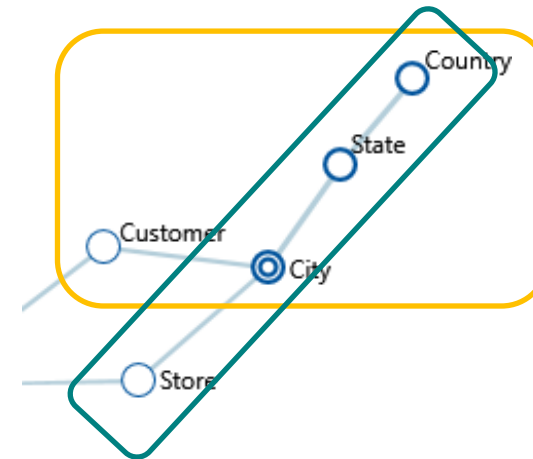
Gerarchie



- Ogni path dalle foglie alla radice diventa una gerarchia separata e gli attributi comuni vengono duplicati

- Le gerarchie condivise vengono duplicate

S. Country S. State S. City



Green VS Blue

- | In Tableau, il colore verde è associato a campi **continui** mentre il colore blu a quelli **discreti**

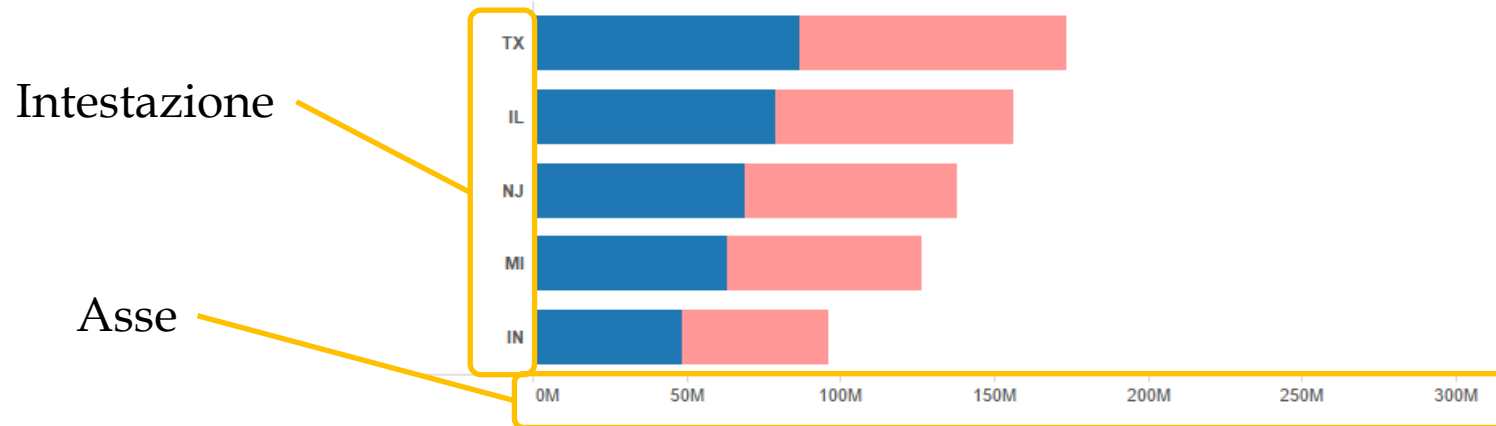


- | Spesso (ma non sempre) le misure sono campi continui, mentre le dimensioni sono campi discreti
- | Campi continui e discreti producono effetti diversi
 - * Quando vengono impostati in righe e colonne
 - * Quando vengono utilizzati in un filtro
 - * Quando vengono associati a colori

Green VS Blue (Righe e Colonne)

I Quando vengono assegnati a righe e colonne

- * Un campo *discreto* produce un'*intestazione*
- * Un campo *continuo* produce un *asse*



Green VS Blue (Filtri)

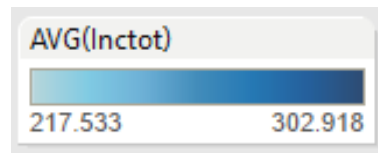
- | Con un campo continuo è possibile specificare dei **range**
 - * Sui valori al livello più **dettagliato**
 - * Oppure su particolari **aggregazioni**

#	All values
#	Sum
#	Average
#	Median
#	Count
#	Count (Distinct)
#	Minimum
#	Maximum
#	Standard deviation
#	Standard deviation (Population)
#	Variance
#	Variance (Population)

- | Con un campo discreto è possibile anche selezionare valori specifici (i.e., uno ad uno)

Green VS Blue (Colori)

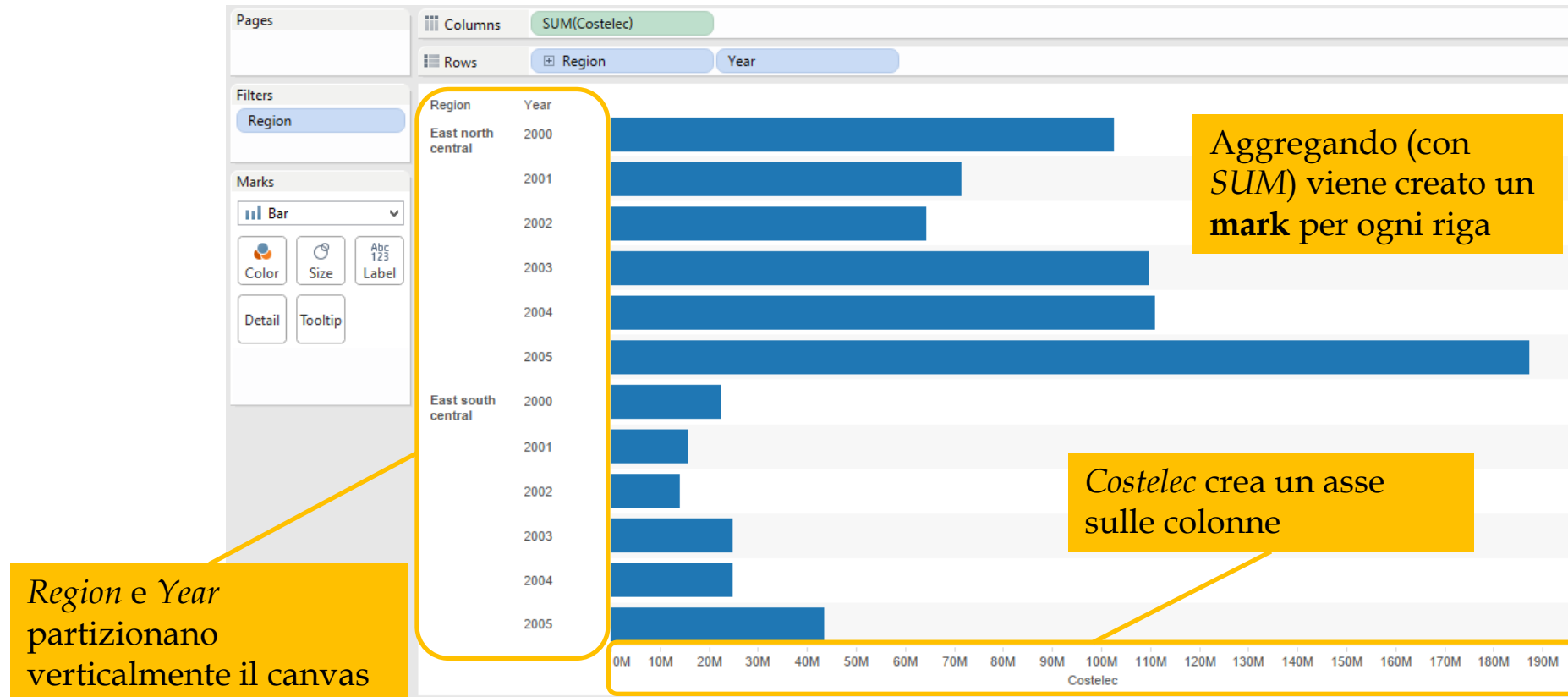
- Ad un campo continuo viene associato un insieme di colori **sequenziali** (e.g. gradazioni più chiare per valori bassi e gradazioni più scure per valori alti)



- Con un campo discreto è possibile assegnare un colore diverso (non necessariamente correlato agli altri) per ogni valore distinto



Canvas



Canvas (**P**ane e **C**ell)

Pages

Columns: MRN, Racegroup

Rows: Region, State, Year

Filters: State, Region, Racegroup, MRN, Year

Marks: SUM(Costelec)

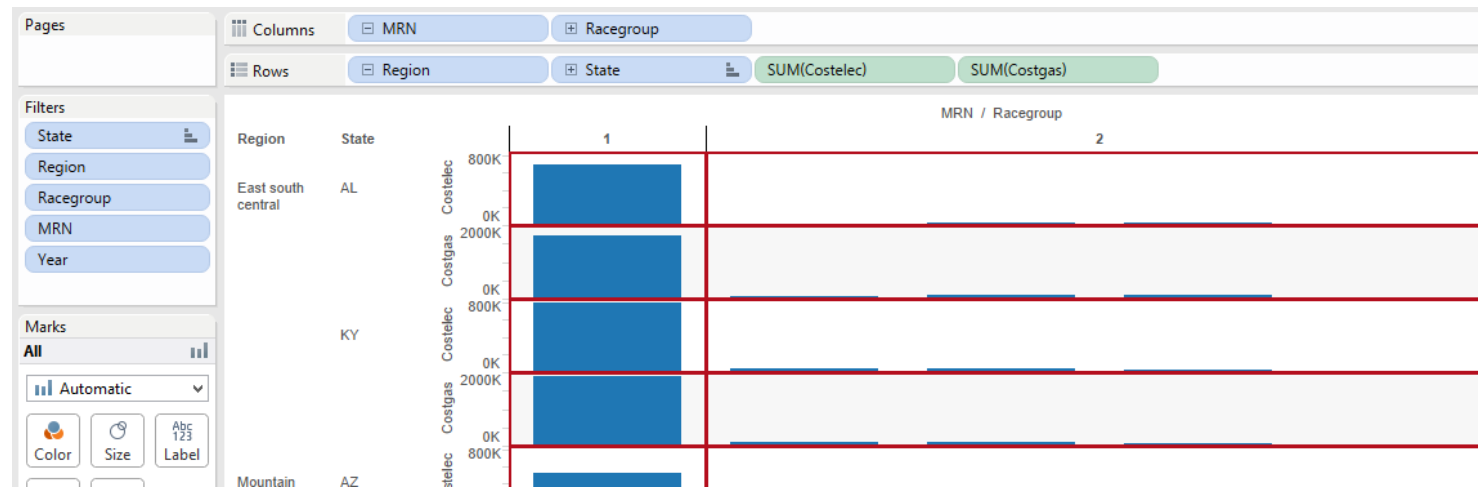
Region	State	Year	Black and AIAN	Black and Asian	Black and Pacific Islander	Asian and Pacific Islander and 'other race' write-in	Black and American Indian/Alaska Native and Asian	Black and Asian and Pacific Islander
East south central	AL	2003	16.687	10.605			1.584	1.092
		2004	29.097	3.396		3.036		3.048
		2005	49.598	28.488				14.881
KY		2003	4.740	1.704	2.076			3.528
			21.669	3.528				
			38.509	18.069	852		6.600	4.884

Le *cell* sono definite dalle combinazioni di valori degli ultimi campi nelle righe e nelle colonne (*Racegroup* e *Year*)

I *pane* sono definiti dalle combinazioni di valori dei penultimi campi nelle righe e nelle colonne (*MRN* e *State*)

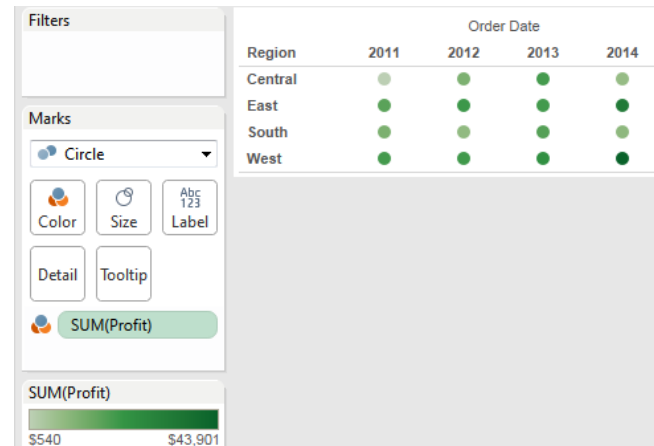
Canvas (**Pane** e **Cell**) (2)

- | In caso di campi continui
 - * Dato un valore distinto dell'ultimo campo discreto, viene definito un pane per ogni campo continuo (i campi continui sono sempre posizionati per ultimi!)
 - * Una cell è un punto nello spazio definito dagli assi



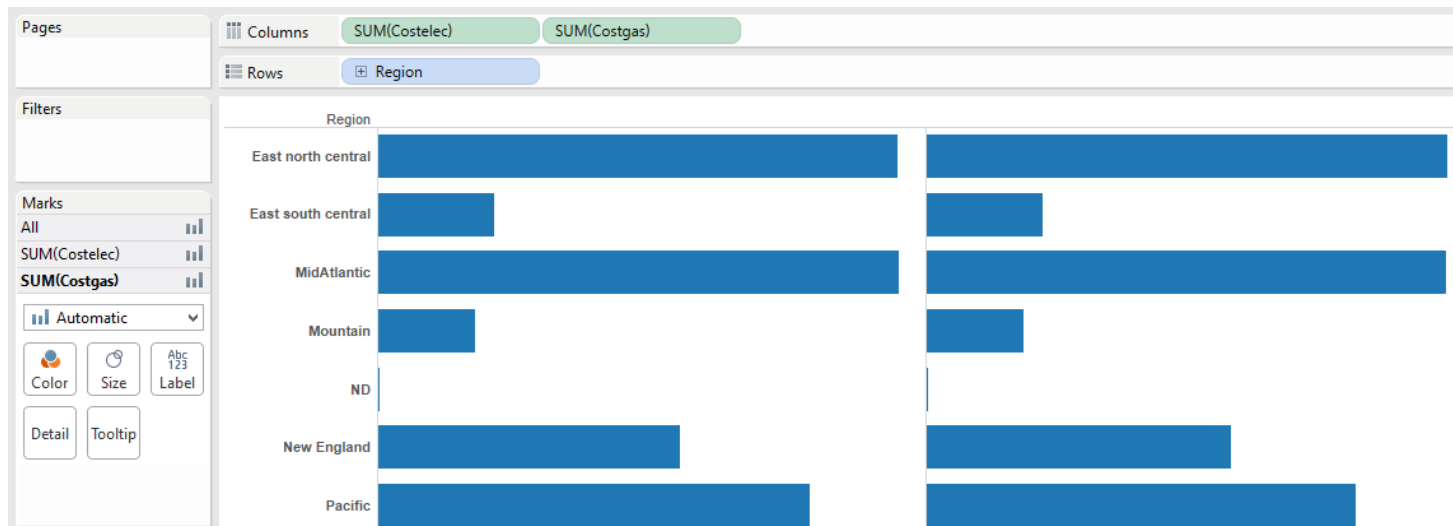
Mark

- | I dati vengono visualizzati all'interno del canvas tramite *mark*
- | Esistono diversi tipi di mark (*bar*, *line*, *text*, etc.)
 - * Ogni mark possiede diverse *proprietà* (*colour*, *size*, *label*, etc.)
 - * A ciascuna proprietà può essere associato un campo con effetti diversi in base alla proprietà e al tipo di campo (continuo o discreto)



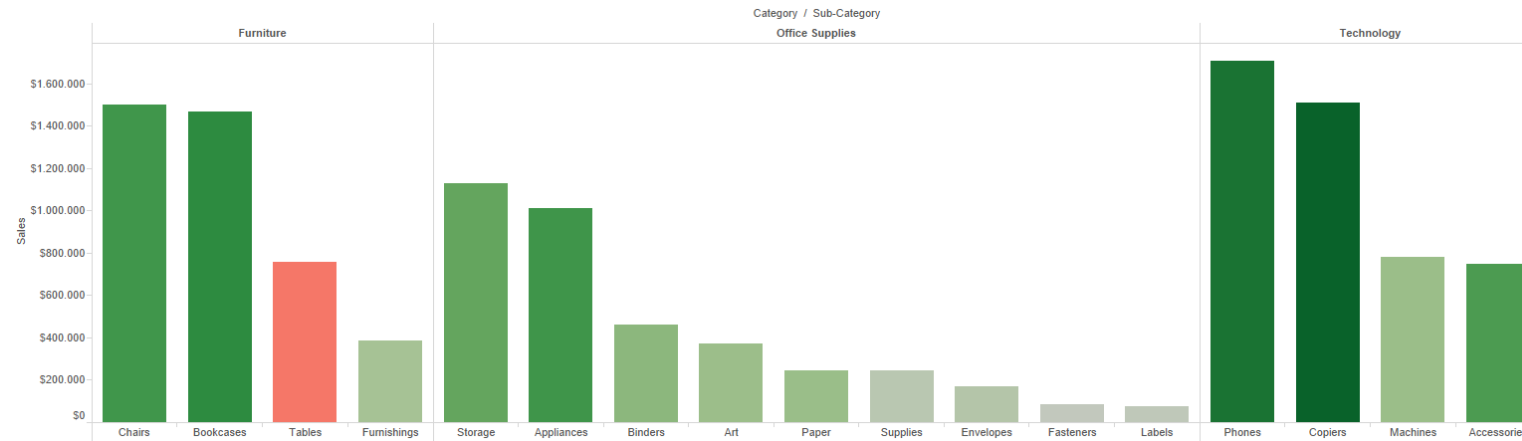
Mark (2)

- | Una cell può contenere da zero a più mark
 - * Ad esempio nel caso in cui vengano visualizzate più misure
- | In uno stesso canvas possono essere presenti più tipi di mark



Ordinamento

- | Esistono tre tipologie di ordinamento
 - * *Manual*: l'ordinamento è fissato manualmente dall'utente
 - * *Computed*: l'ordinamento si basa su un calcolo (eg. la somma di una certa misura)
 - * *Data Source Order*: l'ordinamento è lo stesso della sorgente dati
- | Gli ordinamenti rispettano le gerarchie (ci sono workaround)

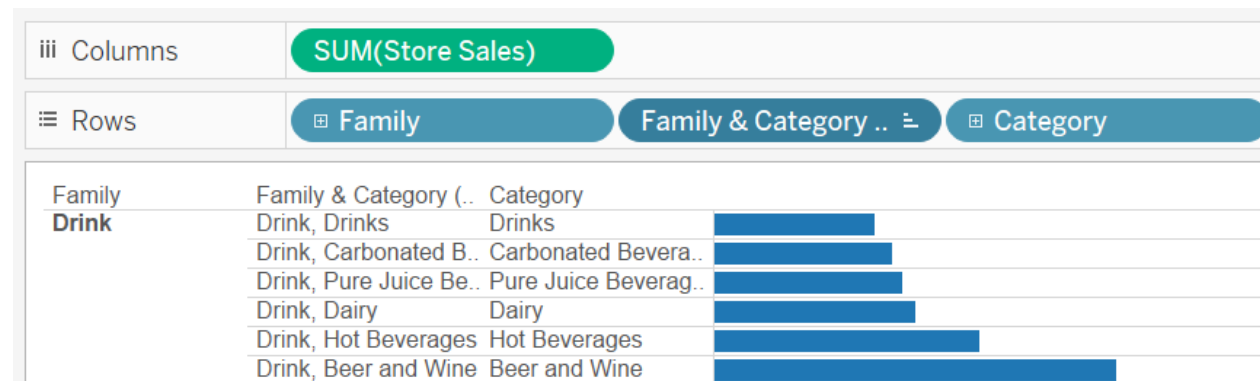


Ordinamento (2)

- Un comportamento inaspettato si presenta quando si cerca di ordinare un campo a destra di un altro che non lo determina funzionalmente

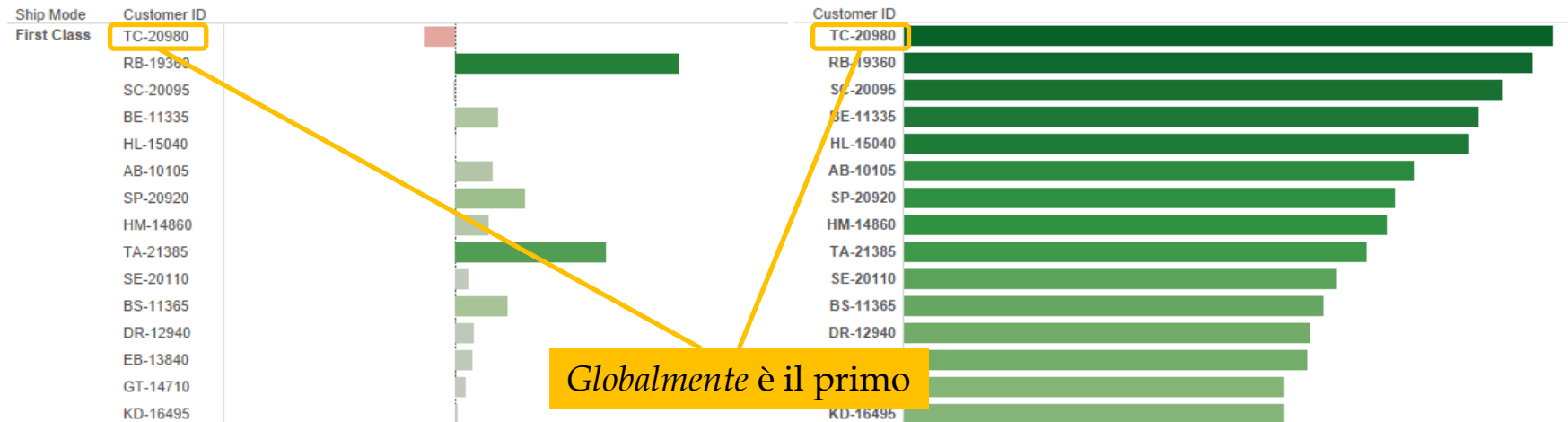


- Workaround:** utilizzare un campo combinato



Ordinamento (3)

- Un comportamento inaspettato si presenta quando si cerca di ordinare un campo a destra di un altro che non lo determina funzionalmente



- Workaround*: utilizzare un campo combinato

View Data

- | Tramite l'opzione *View Data* è possibile visualizzare l'insieme record (i.e., i dati a granularità più fine) utilizzati per calcolare un determinato mark
- | View Data può essere considerato come una versione light dell'operazione *Drill Through*
- | Particolarmente utile per test e debug quando si creano visualizzazioni complesse

View Data (2)

The screenshot shows a map application interface. On the left, there is a sidebar with a search icon, zoom in (+), zoom out (-), and a pin icon. The main map area displays a map of the United States with California highlighted in blue. A context menu is open over California, listing various actions: Select All, View Data... (highlighted), Copy, Format..., Edit Locations..., Mark Label, Annotate, Trend Lines, Forecast, Drop Lines, Hide Map Search, Hide View Toolbar, Keep Only (checked), Exclude, Group, and Create Set... (with a magnifying glass icon).

In the foreground, a 'View Data' dialog box is open. It shows a table with 75,000 rows. The table has three columns: City, MRN, and Occupation. The dialog also includes checkboxes for 'Show aliases' and 'Show all fields', and buttons for 'Copy' and 'Export All'. At the bottom, there are tabs for 'Summary' and 'Underlying', and a row count of 75,000 rows.

City	MRN	Occupation
Los Angeles	1	First-Line Supervisors/Managers of Construction Trades and Extraction Workers
Los Angeles	1	Painters, Construction and Maintenance
Los Angeles	1	Carpenters
Los Angeles	1	Construction Laborers
Los Angeles	1	Carpenters
Pomona	1	Electricians
Riverside	1	Electricians
San Diego	1	Carpet, Floor, and Tile Installers and Finishers
Los Angeles	1	Electricians
Los Angeles	1	Carpenters

Show Me

- La palette *Show Me* contiene scorciatoie per produrre visualizzazioni di tipi differenti a partire da un insieme di dimensioni e misure
- Per poter utilizzare una visualizzazione tramite Show Me è necessario rispettare determinati *requisiti* che variano di caso in caso (e.g. per uno scatter plot sono necessari campi continui)
- Alcuni tipi di visualizzazioni sono poco intuitive da costruire manualmente (e.g. mappe e box-plot)



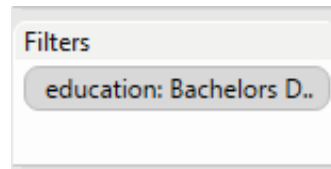
For **scatter plots** try
0 or more dimensions
2 to 4 measures

Filtri

- | È possibile applicare un filtro (i.e., *Slice & Dice*)
 - * A specifici **worksheet**: il filtro è applicato solamente a quegli specifici worksheet
 - * Ad una specifica **sorgente dati**: il filtro verrà implicitamente applicato a tutti i worksheet che estraggono dati da quella sorgente
- | È possibile filtrare
 - * A **livello di record**: la vista è calcolata considerando solamente i record che soddisfano il filtro; ogni filtro è calcolato indipendentemente dagli altri
 - Eg. `Sales > 100.00$`
 - * A **livello di aggregazione**: dopo che la vista è stata calcolata (applicando i filtri a livello di record) vengono escluse le cell per cui almeno un mark non soddisfa il filtro aggregato
 - Eg. `SUM(Sales) > 100.00$`

Filtri: Context Filter

- | I **Context Filter** sono un particolare tipo di filtro che viene applicato *prima* dei normali filtri (gli altri filtri sono dipendenti dal risultato dei context filter)
- | Quando viene creato un context filter Tableau crea una *tabella temporanea* in modo da snellire successivi calcoli
- | I context filter non possono essere però applicati a livello di aggregazione

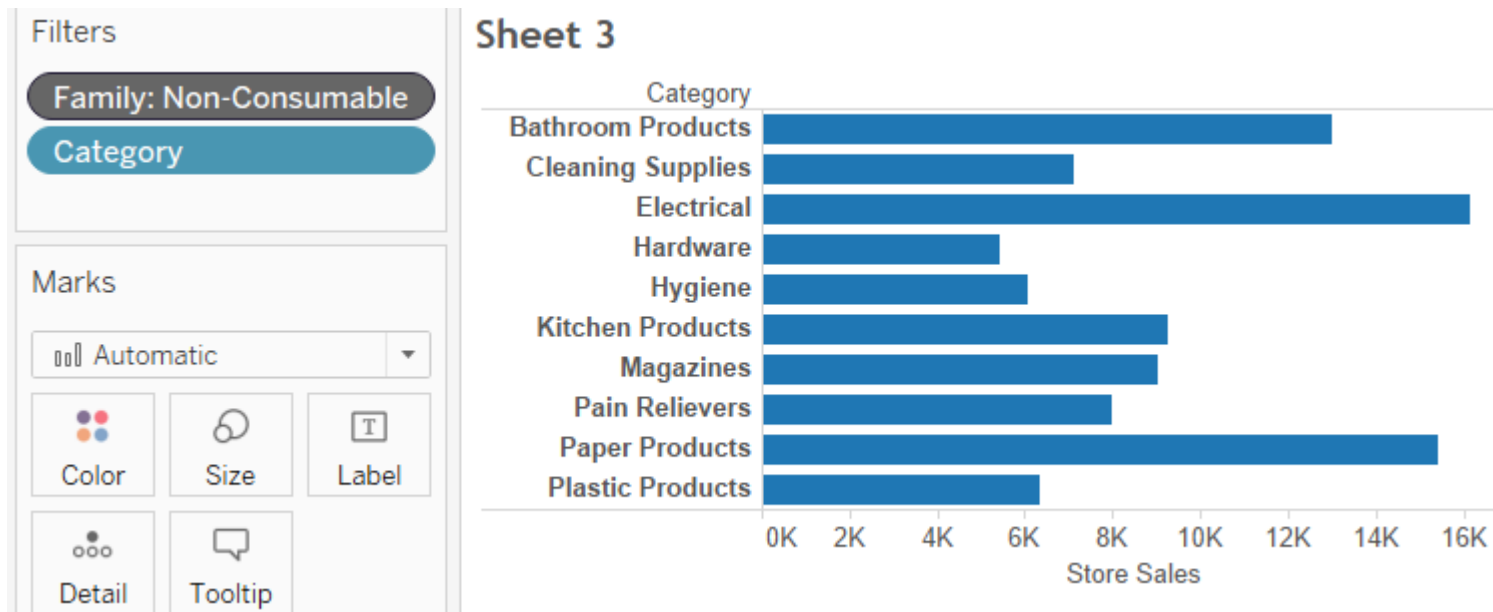
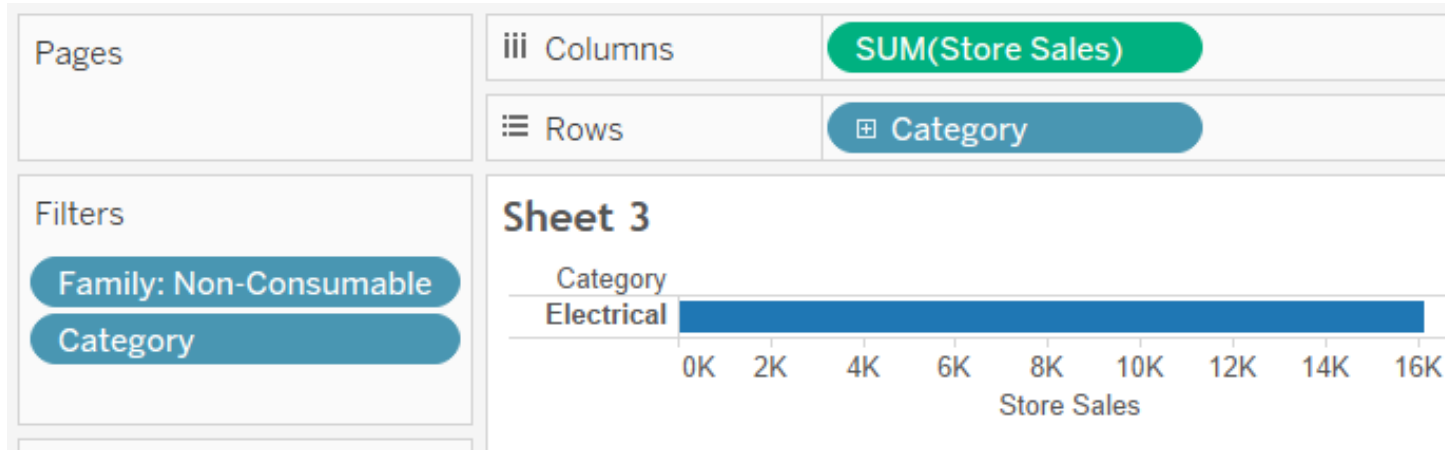


- I context filter sono riconoscibili dal colore grigio (sia per campi continui che discreti)

Top N

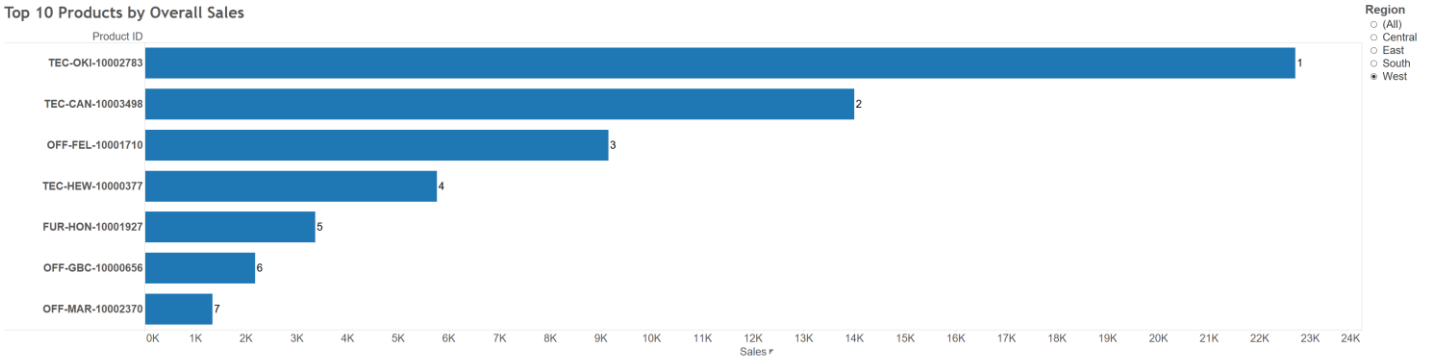
- | È possibile filtrare in modo tale da ottenere solamente i primi (o gli ultimi) N elementi in base ad un determinato ordinamento
 - * E.g. le prime 10 categorie per cui la somma delle vendite è più elevata
- | Attenzione: i filtri top / bottom N vengono applicati **indipendentemente** dagli altri filtri e dalla visualizzazione
 - * E.g. selezionando una famiglia e impostando un filtro top 10 sulle categorie risulterebbe in una visualizzazione in cui sono presenti le categorie che **globalmente** sono tra le top 10!

Top N (2)

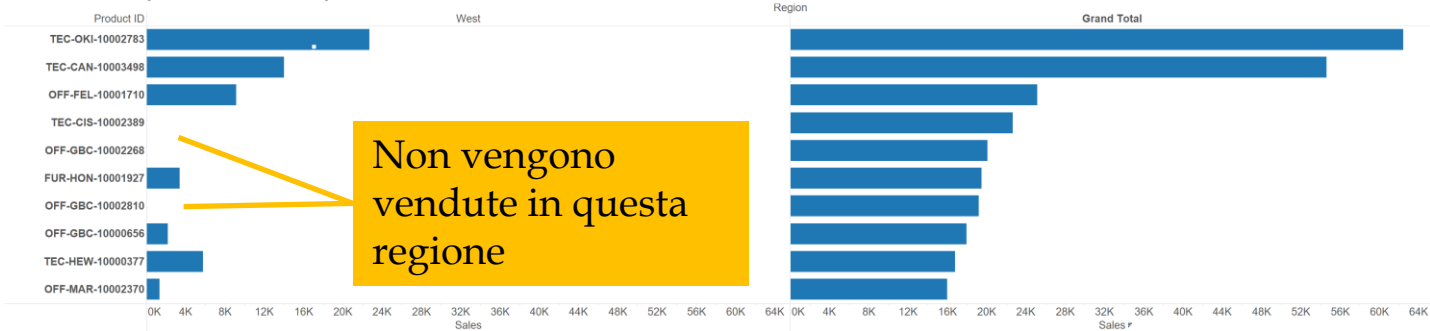


Top N (3)

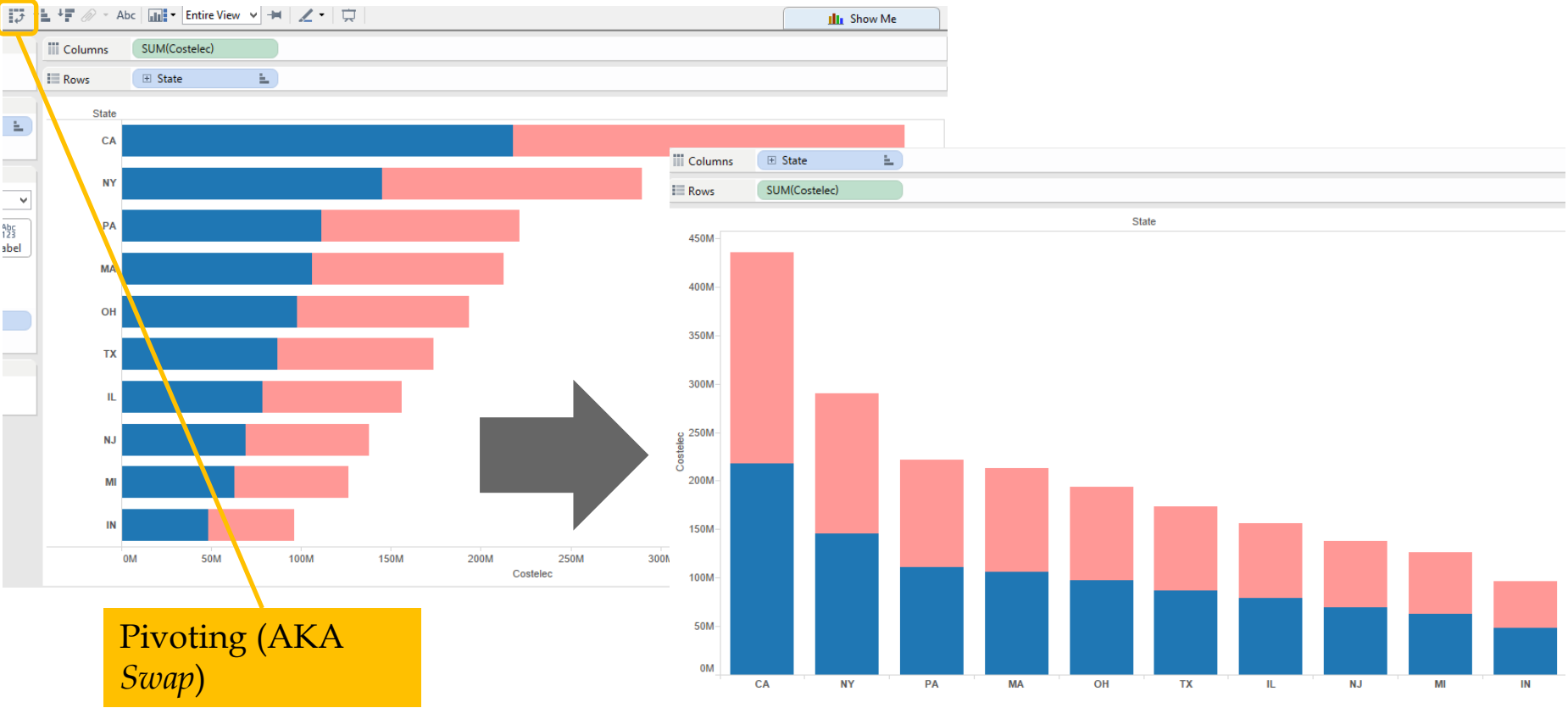
Top 10 Products by Overall Sales



Several of the products in the Top 10 overall aren't sold in the West



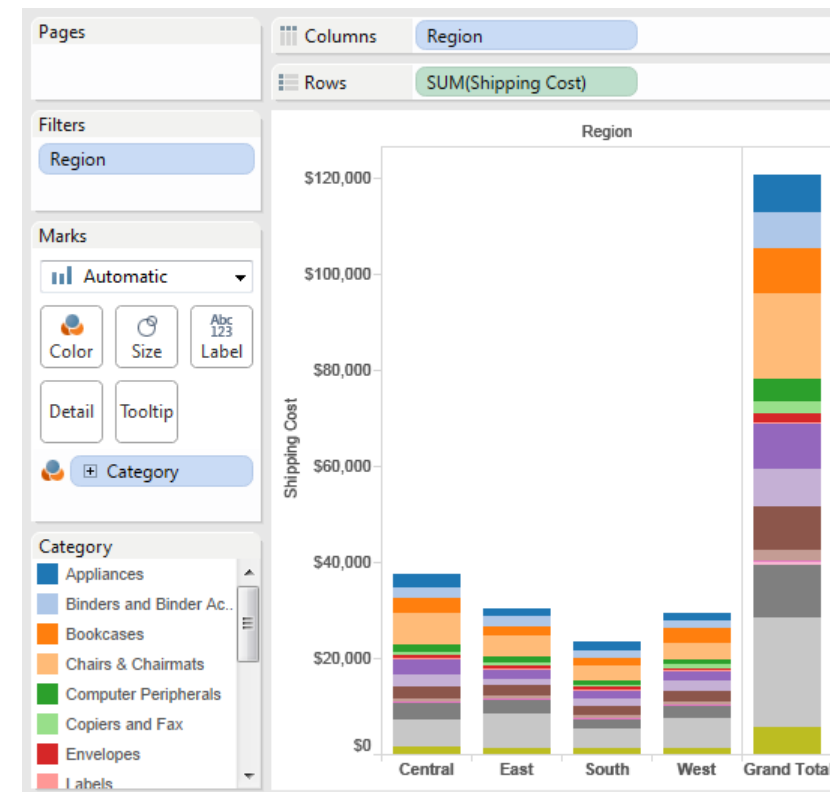
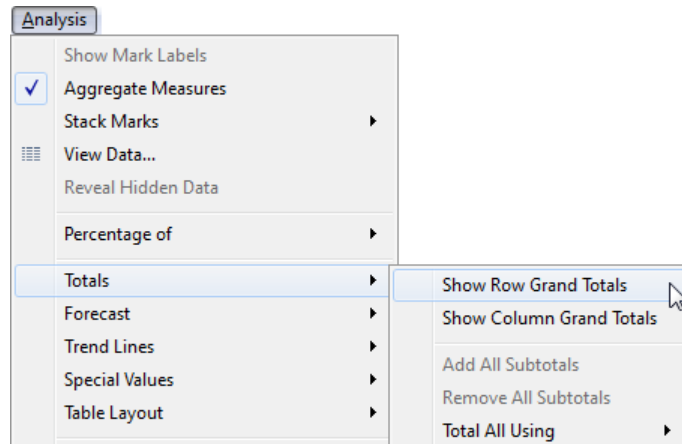
Pivoting



Pivoting (AKA
Swap)

Grand Total e Sub Total

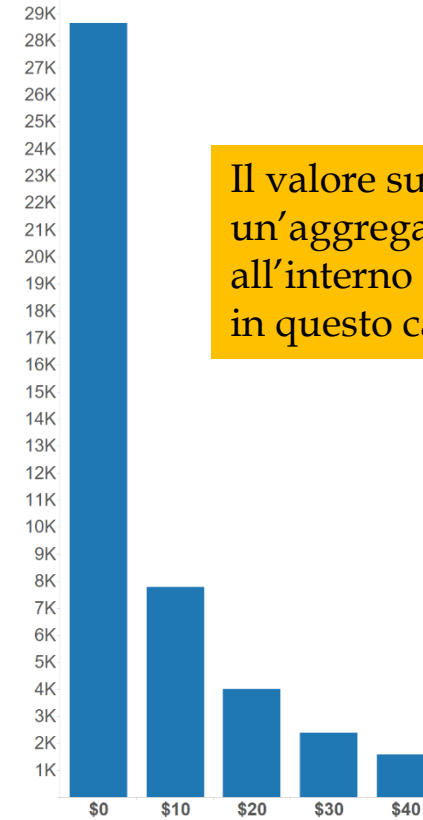
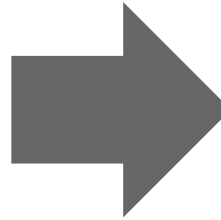
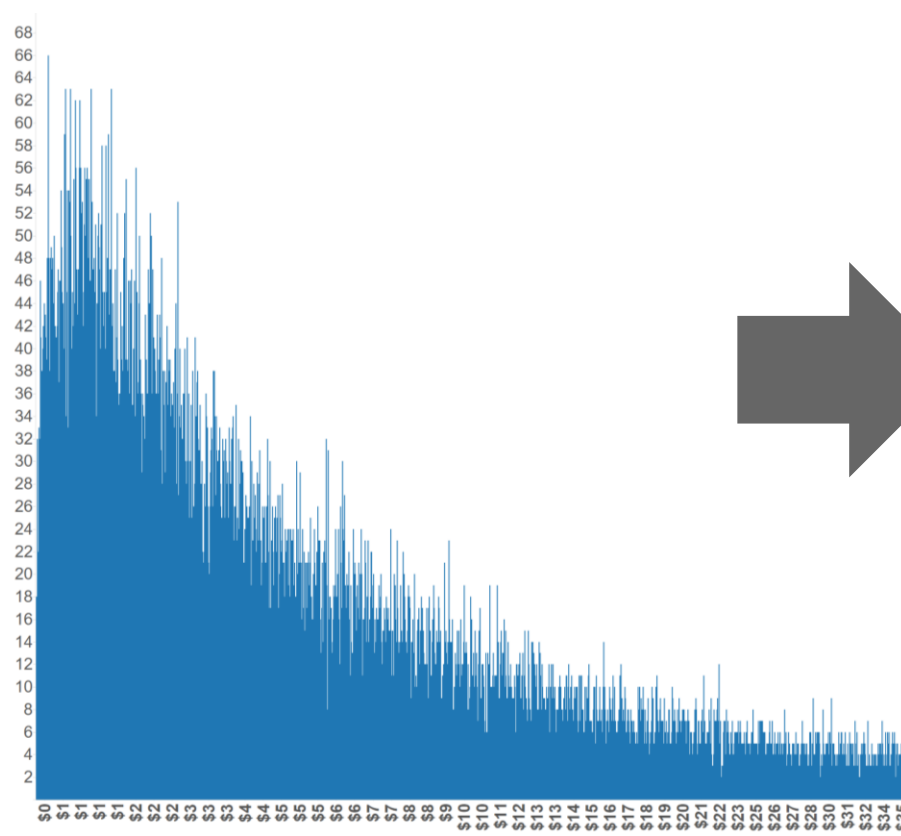
- I totali sono utili per mostrare diversi livelli di aggregazione nella stessa visualizzazione
- Possono essere applicati solamente a campi discreti



Binning

- | Con *binning* si intende la creazione di *bin* (o *bucket*), ovvero *intervalli* numerici che raggruppano valori di una o più variabili
 - * In Tableau gli intervalli sono inclusivi a sinistra ed esclusivi a destra: [*start*, *end*)
- | Rappresentando i dati tramite bin è possibile *discretizzare* variabili continue o comunque di ridurre il numero di valori in caso di variabili discrete
- | Il binning può essere utile per ridurre l'effetto di piccoli scostamenti considerati rumore (e.g. *smoothing*) ed è usato per la creazione di *istogrammi*
- | Le misure vengono aggregate per bin
 - * Eg. la media delle vendite associata ad un bin può essere la media delle vendite degli ordini che ricadono all'interno del bin

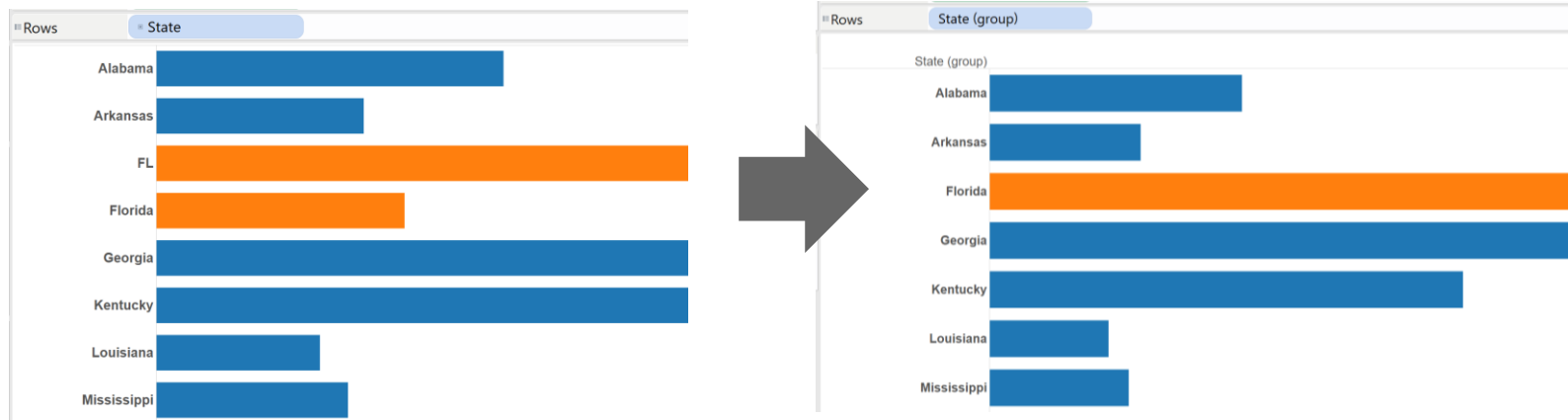
Binning (2)



Il valore sulle ordinate è un'aggregazione all'interno del bin (*sum* in questo caso)

Group

- È possibile creare nuovi campi raggruppando i valori di campi già esistenti; utile ad esempio per
 - * Raggruppare valori che hanno la stessa semantica ma sono etichettati diversamente
 - * Ottenere nuovi raggruppamenti intermedi senza modificare la sorgente dati

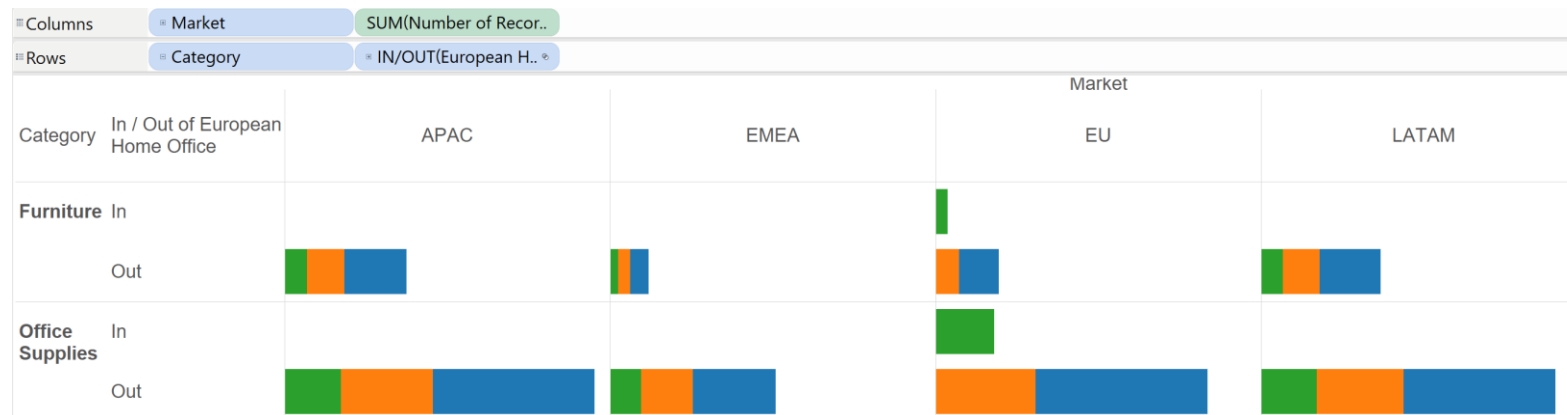


Set

- | In Tableau, un **set** è un insieme di combinazioni di valori dimensionali (i.e., non contiene misure). Esistono due tipologie di set
 - * **Constant**: gli elementi vengono selezionati manualmente e rimangono gli stessi anche se i dati variano
 - * **Computed**: gli elementi vengono selezionati in base ad una formula e si aggiornano automaticamente al variare dei dati
- | I set non possono essere creati a partire da visualizzazioni in cui non ci sono misure aggregate
- | I set possono essere utilizzati in vari modi: come filtri, per applicare diversi colori, come livelli in gerarchie, all'interno di formule, ...
- | I set possono essere **combinati** tra loro per ottenere nuovi set
 - * Eg. creare un nuovo set come unione di altri due

Set (2)

- Se usati come filtri, i set escludono tutti i valori non appartenenti al set
- Se assegnati come colore o utilizzati come dimensione, i set dividono i mark o il canvas in elementi che appartengono al set (*In*) ed elementi che non appartengono al set (*Out*)



Calculated Field

- | Tramite *Calculated Field* è possibile definire nuovi campi dinamicamente senza modificare la sorgente dati
- | Un calculated field è definito da una formula che può utilizzare campi già esistenti e numerose funzioni (logiche, numeriche, su stringhe, su date, etc.)
 - * Eg. Il campo *Profit* può essere definito come *Sales* - *Cost*
- | Un calculated field può essere definito a diversi livelli di granularità
 - * *Line Granularity*: il campo viene calcolato tupla per tupla
 - * *Aggregated Granularity*: il campo viene calcolato su aggregazioni di campi
- | Un calculated field può essere (generalmente) utilizzato come qualunque altro campo, ad eccezione dei campi con granularità aggregated, per i quali è possibile filtrare solo se sono campi continui

Calculated Field: Sintassi

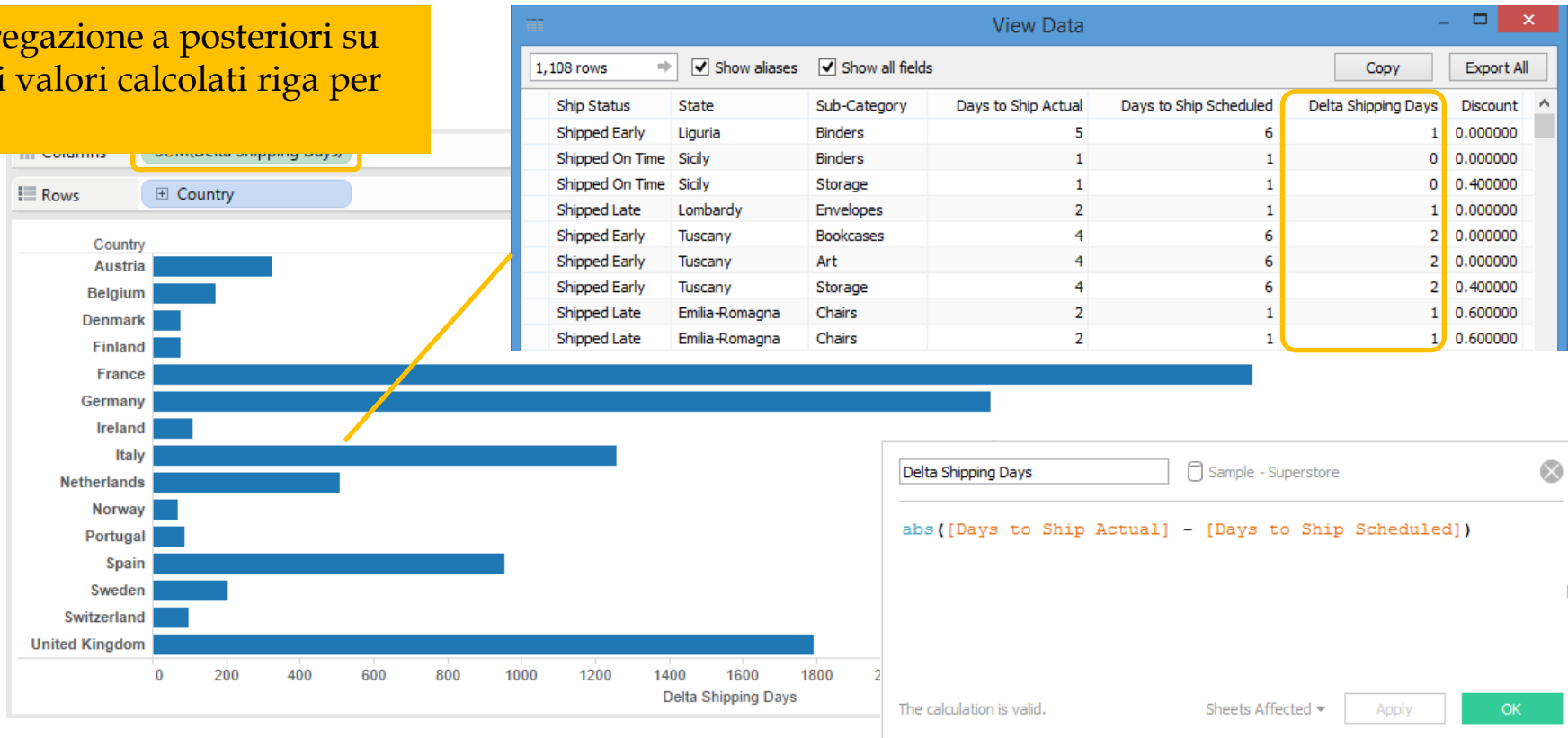
- | Per utilizzare un campo esistente in una formula basta scriverne il nome all'interno di parentesi quadre, eg. `[Sales]`
- | È possibile utilizzare **costrutti condizionali**

```
if [Profit] > 0 then
    'Proficuo'
else
    'Non proficuo'
end
```
- | Per utilizzare una funzione è necessario indicarne il nome e inserire tra parentesi tonde gli argomenti di input separati da virgola

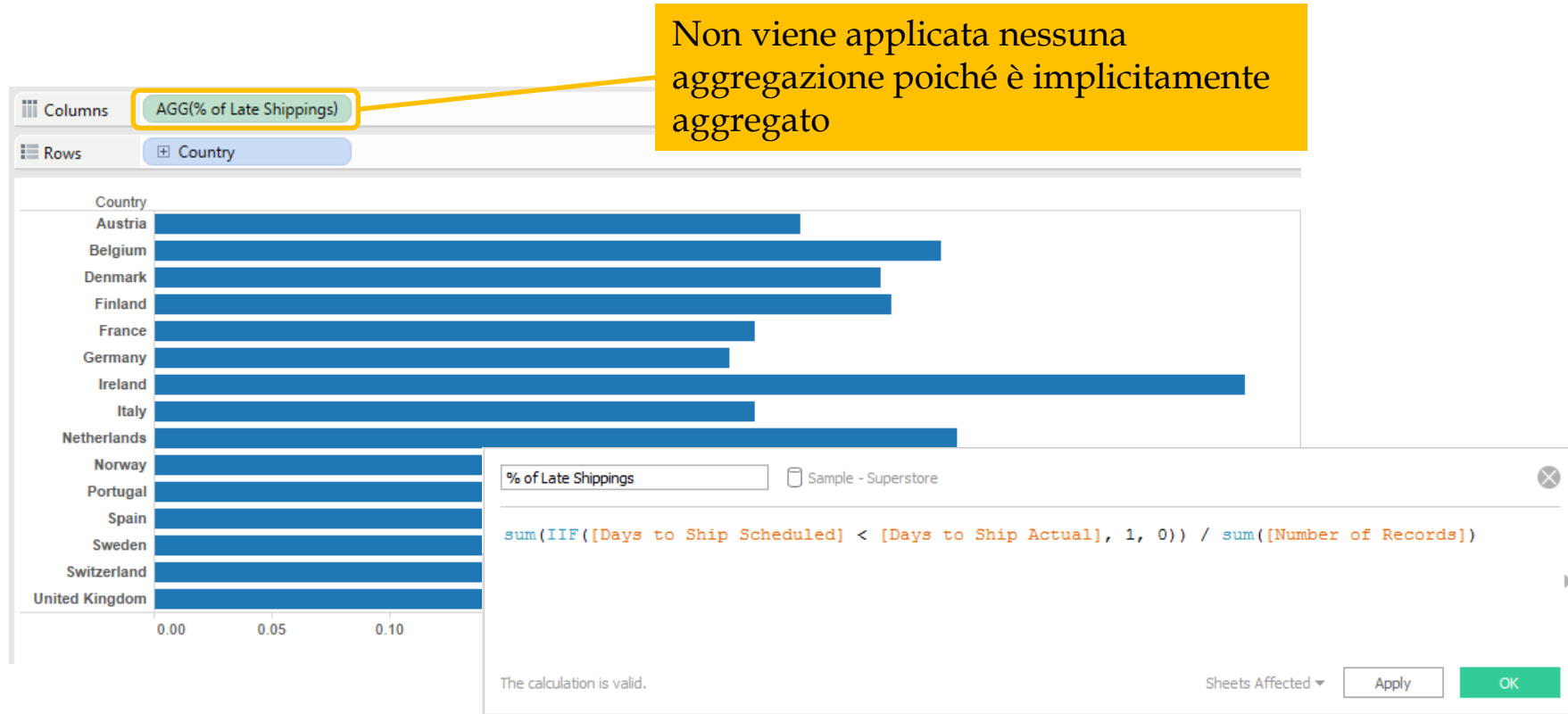
```
* MAX([Sales], [Cost])
```

Calculated Field: Line Granularity

Aggregazione a posteriori su tutti i valori calcolati riga per riga

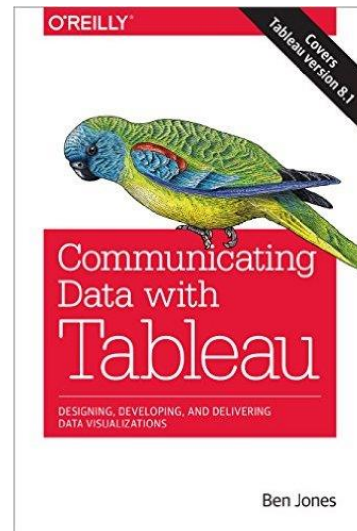


Calculated Field: Aggregated Granularity



Riferimenti

- | Tutorial: <http://www.tableau.com/learn/training>
- | Knowledge Base: <http://kb.tableau.com/>



ESERCIZI – PRIMA PARTE

Esercizio 1

- | Visualizzare tramite un grafico a barre la somma delle *Store Sales* per ogni *S. State*
 - * Qual è lo stato con le vendite più elevate?
- | Effettuare un drill-down per visualizzare le vendite a livello di *S. City*
 - * Esistono città con vendite molto inferiori rispetto alle altre?
- | Quanti sono i negozi (*Store*) presenti in ogni *S. State*?
E in ogni *S. City*?
 - * **Tip:** Utilizzare l'aggregazione *COUNTD*
 - * È possibile imputare le basse vendite in alcune città al numero di negozi?
- | Visualizzare le vendite a livello di *S. City* (come fatto in precedenza) e associare alla proprietà *color* numero di negozi distinti (*COUNTD(store)*)

Esercizio 2

- | Data l'ultima visualizzazione creata in Esercizio 1, associare il campo *S. Type* alla proprietà *color*
 - * Quale pattern interessante è possibile notare?
- | Visualizzare le *vendite (SUM)* per ogni *S. Type*
 - * Quale discrepanza è possibile notare rispetto al grafico precedente?
- | Associare il *numero di negozi (Store)* alla proprietà *color* e alla proprietà *label*
 - * Da cosa è causata la discrepanza tra le due visualizzazioni precedenti?

Esercizio 3

- | Visualizzare tramite un grafico a linee l'andamento **mensile** delle **vendite**
 - * Quale pattern è presente?
- | Dividere il grafico precedente per **S. State** (un asse per ogni stato)
 - * Il pattern precedente è presente in ogni stato?
 - * **Tip:** di default gli assi hanno tutti lo stesso range: su un asse qualsiasi, click destro > *Edit Axis* > Selezionare *Independent axis...*
- | Dato il grafico precedente, visualizzare quanto impattano le varie **Family** sul totale delle vendite mantenendo la visualizzazione del trend mensile
 - * Quale può essere una buona visualizzazione?
 - * **Tip:** associare ogni **Family** ad una proprietà dei mark ed eventualmente cambiare tipologia di mark
 - * **Tip:** è possibile cambiare il tipo di mark dal menu a tendina nel pannello *Marks*

Esercizio 4

- | Visualizzare tutti i **negozi** e ordinarli in ordine decrescente per **somma delle vendite**
- | Aggiungere alla visualizzazione precedente l'attributo dimensionale **Type**
- | Data la visualizzazione precedente, ordinare i negozi in ordine decrescente per **numero di clienti (Customer)**
 - * Per chiarezza, associare il numero di clienti alla proprietà **color**
- | Visualizzare in ordine decrescente la **somma delle vendite** per **Type** e **S. State**
 - * All'interno di ogni **Type** alcuni campi non sono ordinati correttamente...
 - * **Tip:** per creare un *campo combinato* è necessario selezionare (dal menu delle dimensioni) due campi, *click destro > Create > Combined Field*

Esercizio 5

- | Visualizzare le vendite per *Occupation* (dimensione *Customer*) escludendo tutte le tuple con un importo minore di 5
 - * *Tip*: applicare un filtro sul campo *Store Sales*
- | Data la visualizzazione precedente (mantenere anche il filtro), applicare un filtro che scarti tutte le *Occupation* per cui la somma delle vendite è inferiore a 80K
 - * Sono ancora presenti alcune *Occupation* con vendite inferiori a 80K... Come è possibile spiegare questo comportamento di Tableau?
 - * Fare in modo che il filtro aggregato (i.e., sulla somma delle vendite) venga applicato *dopo* quello sulle singole tuple
 - * *Tip*: per trasformare un filtro in un *context filter*, click destro > *Apply to Context*

ESERCIZI - SECONDA PARTE

Esercizio 6

- | Visualizzare i **dieci clienti** (*Customer*) con la più alta **somma di vendite**
 - * **Tip:** un filtro **Top N** può essere applicato trascinando un campo nel pannello *Filters* ed utilizzando l'apposita tab *Top*
- | Data la visualizzazione al punto precedente, aggiungere il campo *Occupation*
 - * Tutte le occupazioni non hanno dieci clienti? Perché?
- | Data la visualizzazione al punto precedente, filtrare per *Occupation* e selezionare il valore *Professional*
 - * Quanti clienti sono visualizzati? È possibile fare in modo che vengano visualizzati i Top N clienti relativamente a *Professional*?
 - * **Tip:** vedi Esercizio 5

Esercizio 7

- | Visualizzare la distribuzione delle tuple per *Store Sales*
 - * *Tip*: uno strumento molto utile per analizzare distribuzioni è l'*istogramma* (vedi *Histogram* nel pannello *Show Me*)
 - * Che forma ha la distribuzione risultante?
 - * Quali vendite contiene il bin etichettato 0 (utilizzare *View Data*)? Contiene solo le vendite con importo = 0?
- | Senza utilizzare il pannello *Show Me*, visualizzare un grafico a barre con *Store Sales* sulle colonne raggruppati in bin di dimensione 2, e con la somma di *Store Cost* sulle righe
 - * *Tip*: dal pannello *Measures*, click dx sul campo *Store Sales* > *Create* > *Bins*
 - * Non sarebbe intuitivo aspettarsi che all'aumentare delle vendite aumentino anche i costi? Perché questa visualizzazione è fuorviante?
 - * Modificare il grafico in modo da mostrare la correlazione tra *Store Cost* e *Store Sales*

Esercizio 8

- | Creare un *set* con i **Top 500 clienti** per **somma di Store Sales**
 - * **Tip:** per creare un set, click destro sul campo *Customer* > *Create* > *Set*
- | Posizionare la **somma delle vendite** sulle colonne e il **set creato al punto precedente** sulle righe
 - * Cosa rappresenta la visualizzazione risultante?
- | Modificare la visualizzazione precedente spostando il **set** sulla proprietà **color** e aggiungendo **S. Country** sulle righe
 - * Data questa visualizzazione è possibile vedere che i top 500 contribuiscono in modo considerevole al totale?
 - * Dopo aver effettuato drill-down sulla gerarchia *S. Location*, è possibile notare una distribuzione omogenea o eterogenea delle vendite dei top 500?

Esercizio 9

- | Visualizzare l'andamento dei profitti ($\text{Profitto} = \text{Vendite} - \text{Costi}$) mese per mese per ogni *Type*
 - * *Tip*: per creare un *Calculated Field*, dal menu principale (in alto), *Analysis > Create Calculated Field...*
- | Per ogni utente, calcolare l'età e visualizzarne l'istogramma considerando bin di dimensione pari a cinque
 - * *Tip*: la funzione *DATEDIFF* restituisce la differenza di due date (vedi descrizione data cliccando sul nome della funzione durante la creazione del campo)
 - * Esiste qualche gruppo di età che si comporta in modo differente dagli altri?
- | Data la visualizzazione precedente, sostituire il numero di vendite con il numero di vendite rapportato per il numero di clienti
 - * Il pattern al punto precedente è ancora così evidente? Perché?