DATA & AI LABORATORY

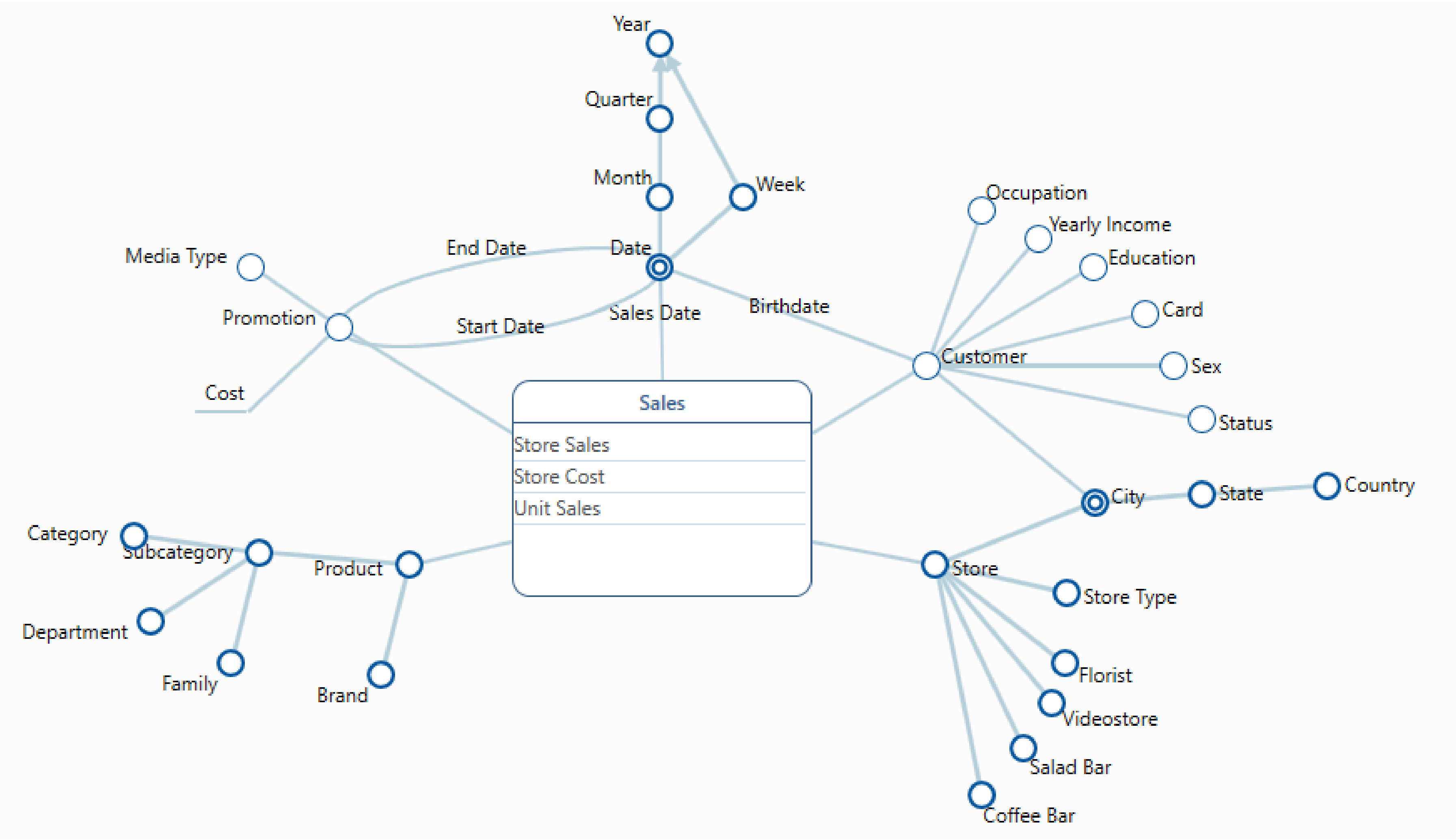
(Big) Data and Artificial Intelligence

Enrico Gallinucci

28/09/2024

BOLOGNA BUSINESS SCHOOL
Alma Mater Studiorum Università di Bologna

DFM – Foodmart



Exercise 1

- Use a bar chart to plot the total sum of *STORE_SALES* for each *STORE_STATE*
  - Which one is the state with the highest sales?
- Apply a drill-down operation to show the sales at the *STORE_CITY* level
  - Are there cities whose sales are much lower than the others'?
- How many stores are there in each *STORE_STATE*? In each *STORE_CITY*?
  - Color the bars based on the *Count(Distinct)* summarization function over the *STORE_NAME* attribute
  - Would it be reasonable to say that cities with fewer stores also have lower total sales?

Exercise 2

- Use a bar chart to plot the total sum of *STORE_SALES* by *STORE_CITY* and assign the *STORE_TYPE* to the Legend property
  - Can you notice any interesting pattern?
- Use a bar chart to plot the total sum of *STORE_SALES* by *STORE_TYPE*
  - Assign the number of stores to the color property
  - Is the result surprising/expected?

Exercise 3

- Use a line chart to plot the monthly sales trend
  - Any interesting pattern?
- Split the previous chart by *STORE_STATE*
  - Put the STORE_STATE in the Legend
  - Does the previous pattern hold for each state?
- Visualise the impact of each *STORE_FAMILY* on the total sales while still showing the monthly trends
  - Use a Stacked area chart, where the *STORE_STATE* is in the Small multiples and the *PRODUCT_FAMILY* in the Legend

Exercise 4

- Analyze sales by *STORE_TYPE* (sorted by descending order)
- Drill-down to the stores
- Add the number of customers
  - Use the *Count(Distinct)* summarization function
  - In case of wrong calculation (i.e., if you get the same value in all rows):
    - Go back to the Model
    - Double-click the relationships between CUSTOMER and SALES
    - Set the *Cross filter direction* to *Both*
- Add the average sales per customer
  - Create a new measure, calculated by dividing the sum of store sales by the count of distinct customers

Exercise 5

- Create a table to visualize the sales for each *OCCUPATION* (*Customer* dimension)
- Exclude (i.e., filter out) the tuples where the value of *STORE_SALES* is lower than 5
- Apply another filter (in addition to the previous one) to exclude all occupations where the total sales is lower than 80K

Exercise 6

- Create a table to visualize with the top ten customers by total sales
  - Show both *CUSTOMER_ID* and *FULLNAME*
- Add the *Occupation* field
- Turn it into a matrix (without the *FULLNAME*)
- Add a measure on the Customer table calculating a ranking of customers
  - First, declare a new measure simply calculating the sum of *STORE_SALES*
  - Then, declare a new measure calculating the RANKX, where
    - The 1st parameter is the attribute that we want to order, i.e., the *CUSTOMER_ID*
    - The 2nd parameter is the measure to be used for ordering, i.e., the one declared above
- Take the first ten customers for each occupation by filtering on the rank

Exercise 7

- Create a histogram of StoreSales
  - Right-click on *STORE_SALES* > New group > Create bins of size 2
  - Create a bar chart showing the count of records for each bin
- Use the same binning to plot a bar chart with the average *STORE_COST* for each bin
  - Do you see a correlation in the data?
- Plot the same result as a scatter chart
  - Find the chart in the list of visuals
  - Put *STORE_COST* and *STORE_SALES* on X and Y axis, respectively (without summarizing)

Exercise 8

- Create a new column calculating the profits
  - *PROFIT = STORE_SALES - STORE_COST*
- Create a line chart showing the monthly trend of profits, sales, and costs

Exercise 9

- Check distribution of profits with respect to different attributes
  - Try some combinations of attributes (e.g., with the stacked bar chart)
    - E.g., check the distribution of profits with respect to different combinations of occupation and gender
  - Is there any categorical values that sticks out? Or are profits mainly driven by the number of customers?
- Create bins where necessary (e.g., population)
  - Calculate the age of customers from their birthdate
    - A new column must be defined
    - <newColumn> = DATEDIFF(<date1>, <date2>, YEAR)
  - Calculate bins of customer ages and check the number of customers in each bin and distribution of profits