CSI 4142 - Fundamentals of Data Science

Project Phase 1: Conceptual Design – Dimensional Model

Due Date: February 8, 2024, 11:59pm

Group 33: Tanner Guiddings, Coralie Ostertag, and Dana Shayakhmetova

1. The Grain

The amount of money a customer spends on various products, while engaging with specific marketing campaigns, making web visits and complaints, which results in a response outcome and recency measure, thus affecting contact costs and revenues.

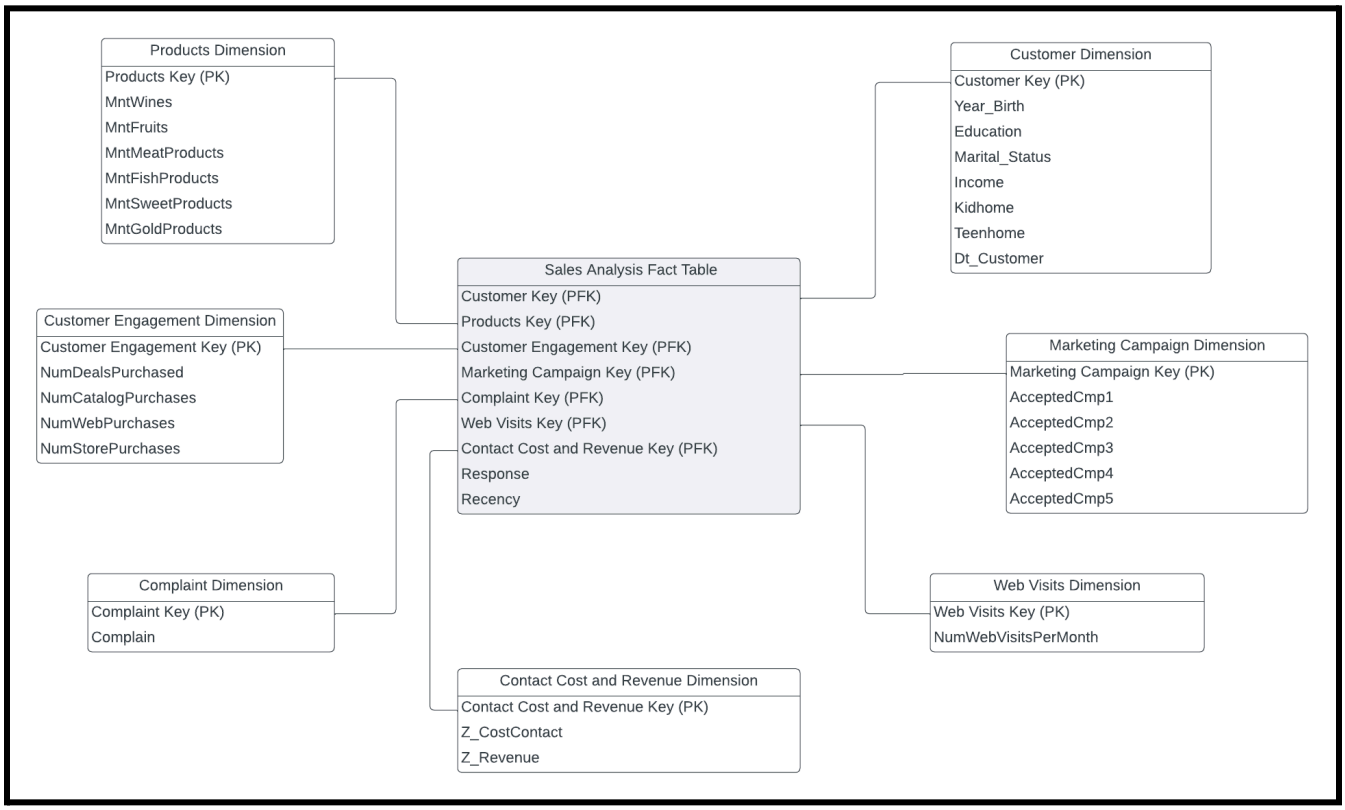

Figure 1: Dimension Modelling Diagram of Our Data Set

2. Dimensions and Attributes

Detail all the dimensions and dimensional attributes. You should list the domains and show sample values. (e.g., Age: integer, minimum = 0 and maximum= 130, Sample value = 35).

Dimension	Variable	Description	Variable Summary
Customer	Year_Birth	The birth year of the individual. (Integer)	Min: 1893 1st Qu: 1959 Median: 1970 Mean: 1969 3rd Qu: 1977 Max: 1996 Sample value: 1969
Customer	Education	The highest level of education attained by the individual. (String)	Graduation, PhD, Master, Basic, 2n Cycle Sample value: PhD
Customer	Marital_Status	The marital status of the individual. (String)	Single, Together, Married, Divorced, Widow, Alone, Absurd, YOLO Sample value: Single
Customer	Income	The annual income of the individual. (Integer)	Min: 1730 1st Qu: 35303 Median: 51382 Mean: 52247 3rd Qu: 68522 Max: 666666 Sample value: 52247
Customer	Kidhome	The number of young children in the household. (Integer)	0,1,2 Sample value: 0
Customer	Teenhome	The number of teenagers in the household. (Integer)	0,1,2 Sample value: 0
Customer	Dt_Customer	The date when the customer was first enrolled or became a part of the company's database. (String)	Min: 2012-07-30 Max: 2014-06-29 Sample value: 2014-06-29
Product	MntWines	Amount spent on wines. (Float)	Min: 0 1st Qu: 23.75 Median: 173.50 Mean: 303.94 3rd Qu: 504.25 Max: 1493 Sample value: 173.5

Product	MntFruits	Amount spent on fruits. (Integer)	Min: 0 1st Qu: 1 Median: 8 Mean: 26.3 3rd Qu: 33 Max: 199 Sample value: 10
Product	MntMeatProducts	Amount spent on meat products. (Integer)	Min: 0 1st Qu: 16 Median: 67 Mean: 166.9 3rd Qu: 232 Max: 1725 Sample value: 10
Product	MntFishProducts	Amount spent on fish products. (Integer)	Min: 0 1st Qu: 3 Median: 12 Mean: 27.53 3rd Qu: 50 Max: 259 Sample value: 10
Product	MntSweetProducts	Amount spent on sweet products. (Integer)	Min: 0 1st Qu: 1 Median: 8 Mean: 27.06 3rd Qu: 33 Max: 263 Sample value: 10
Product	MntGoldProducts	Amount spent on gold products. (Integer)	Min: 0 1st Qu: 9 Median: 24 Mean: 44.02 3rd Qu: 56 Max: 362 Sample value: 10
Customer Engagement	NumDealsPurchases	The number of purchases made with a discount or as part of a deal. (Integer)	Min: 0 1st Qu: 1 Median: 2 Mean: 2.325 3rd Qu: 3 Max: 15 Sample value: 10
Customer Engagement	NumWebPurchases	The number of purchases made through the company's website. (Integer)	Min: 0 1st Qu: 2 Median: 4 Mean: 4.085 3rd Qu: 6

			Max: 27 Sample value: 10
Customer Engagement	NumCatalogPurchases	The number of purchases made through catalogs. (Integer)	Min: 0 1st Qu: 0 Median: 2 Mean: 2.662 3rd Qu: 4 Max: 28 Sample value: 10
Customer Engagement	NumStorePurchases	The number of purchases made in physical stores. (Integer)	Min: 0 1st Qu: 3 Median: 5 Mean: 5.79 3rd Qu: 8 Max: 13 Sample value: 10
Web Visit	NumWebVisitsMonth	The number of visits to the company's website in a month. (Integer)	Min: 0 1st Qu: 3 Median: 6 Mean: 5.317 3rd Qu: 7 Max: 20 Sample value: 10
Marketing Campaign	AcceptedCmp1	Binary indicator (1 or 0) whether the individual accepted the first marketing campaign. (Boolean)	0,1 Sample value: 0
Marketing Campaign	AcceptedCmp2	Binary indicator (1 or 0) whether the individual accepted the second marketing campaign. (Boolean)	0,1 Sample value: 0
Marketing Campaign	AcceptedCmp3	Binary indicator (1 or 0) whether the individual accepted the third marketing campaign. (Boolean)	0,1 Sample value: 0
Marketing Campaign	AcceptedCmp4	Binary indicator (1 or 0) whether the individual accepted the fourth marketing campaign. (Boolean)	0,1 Sample value: 0
Marketing Campaign	AcceptedCmp5	Binary indicator (1 or 0) whether the individual accepted the fifth marketing campaign. (Boolean)	0,1 Sample value: 0
Complaint	Complain	Binary indicator (1 or 0) whether the individual has made a complaint. (Boolean)	0,1 Sample value: 0
Contact Cost and Revenue	Z_CostContact	A constant cost associated with contacting a customer. (Integer, Static)	3 Sample value: 3

Contact Cost and Revenue	Z_Revenue	A constant cost associated with a successful campaign response. (Integer, Static)	11 Sample value: 11

3. Measures/Facts

Detail all the measures/facts. You should list the domains and sample values. (e.g., Age: integer, minimum = 0 and maximum = 130, Sample value = 35).

Attribute	Description	Variable Summary
Response	Binary indicator (1 or 0) whether the individual responded to the marketing campaign. (Boolean)	0,1 Sample value: 0
Recency	The number of days since the last purchase or interaction. (Integer)	Min: 0 1st Qu: 24 Median: 49 Mean: 49.11 3rd Qu: 74 Max: 99 Sample value: 22
Customer Key	Primary key for the Customer Dimension	Random 6 digit number starting with the number **1** Sample value: **1**02240
Products Key	Primary key for the Products Dimension	Random 6 digit number starting with the number **2** Sample value: **2**01122
Customer Engagement Key	Primary key for the Customer Engagement Dimension	Random 6 digit number starting with the number **3** Sample value: **3**04458
Complaint Key	Primary key for the Complaint Dimension	Random 6 digit number starting with the number **4** Sample value: **4**01789
Web Visits Key	Primary key for the Web Visits Dimension	Random 6 digit number starting with the number **5** Sample value: **5**09876
Contact Cost and Revenue Key	Primary key for the Contact Cost and Revenue Dimensions	Random 6 digit number starting with the number **6** Sample value: **6**01564

* Having a unique number at the start of each key helps with readability (example: we know that keys starting with 1 are all Customer Keys).

4. Assumptions

In a dimensional model, the decision to create a separate dimension for certain attributes depends on several factors, including the nature of the data and the analytical requirements. Let's review why we chose to put the "NumWebVisitsMonth" and "Complain" attributes in their own dimensions.

Attribute: NumWebVisitsMonth
Reasoning:

- It represents a count of web visits in a month. We might want to analyze web visit patterns and trends independently. This level of granularity allows us to have a more in-depth understanding of customer behavior on the website. This can be useful for specifically understanding customer engagement with the company's website.
- Web visits may have unique characteristics and patterns that are distinct from other dimensions. Analyzing web visit data in isolation or in combination with other dimensions can provide valuable insights into online customer behavior. We wanted to make sure we set ourselves up to have this option later on in the project.

Attribute: Complain
Reasoning:

- The attribute "Complain" is binary, indicating whether a customer has made a complaint (1) or not (0). If we want to look into complaints during the analysis, having a separate dimension allows for a clear distinction between customers who have complained and those who haven't.
- Essentially, by having a dedicated dimension, it becomes easier to track and analyze customer complaints separately from other aspects of customer behavior.
- Also, if there are future expansions or changes related to customer complaints, having a separate dimension provides flexibility in managing and updating that aspect without affecting other dimensions. This is where we are trying to avoid design mistake number 4 (more on this in section 5).

5. Avoiding Design Mistakes

In this part of the project, we made sure to avoid the following mistakes:
#1: Placing text attributes in the fact table.

- We made sure to place all text attributes in dimension tables rather than in the fact table.
- We used keys for dimension tables to establish the relationships with the fact table.

#2: Limit verbose descriptions to save space.

- We made sure that all the attribute names we used are clear and descriptive to make sure everything is understood and easy to find.
- Additionally, by having the tables in section 2 and 3, users (and our team) can use if as a reference (metadata/documentation) for explanations on the purpose, usage, and context of each dimension, attribute, and measure.

#4: Ignore the need to track changes.

- We saw the example above. Having the “Complain” attribute in a separate dimension allows for changes/modifications to be made without affecting the rest of the structure.

#7: Neglect to declare (and comply with) the grain.

- The first thing we did at the start of this document is declare the grain! ✅
- Throughout the rest of the project, we will make sure we comply.

6. Work Plan

A summary of your team’s work plan, including the times and dates you met, how you divided the work, and how you often meet with the TA.

Time of Action	Description of work
January 18th, 2024	Discussed our initial ideas before meeting with TA.
January 19th, 2024	Had our first meeting with the TA and submitted our plan.
January 19th, 2024	Scheduled a meeting for February 2nd, 2024 to work on the next part of the project.
January 31st, 2024	Started analysis in a shared Colab file.
February 2nd, 2024	Had a meeting to work on the conceptual design.
February 5th, 2024	Had a meeting with TA to address a couple of questions.

As we were working on the conceptual design, we consulted each other through the process of each task. This was to ensure that we all understood what our plans were and so we could synchronize interrelated tasks. The table below identifies which team member was in charge of which task:

Team Member	Task
Dana	Data Mart Schema
	Defining Grain
Coralie	Data Mart Schema
	Assumptions
	Avoiding Design Mistakes
Tanner	Measures/Facts
	Work Plan
	Dimensions and Attributes

7. References

Add a list of additional references you used when creating your model.

Viktor, HL. (n.d). Module 1: Conceptual design through Dimensional Models [PowerPoint slides] https://uottawa.brightspace.com/d2l/le/content/413332/viewContent/5789919/View

Viktor, HL. (n.d). Module 1: More about dimensions [PowerPoint slides] https://uottawa.brightspace.com/d2l/le/content/413332/viewContent/5789919/View

Viktor, HL. (n.d). Module 1: Conceptual design and Grain Conformed [PowerPoint slides] https://uottawa.brightspace.com/d2l/le/content/413332/viewContent/5789919/View

Patel V. (2024). Customer Segmentation: Clustering, Kaggle, https://www.kaggle.com/datasets/vishakhdapat/customer-segmentation-clustering/data