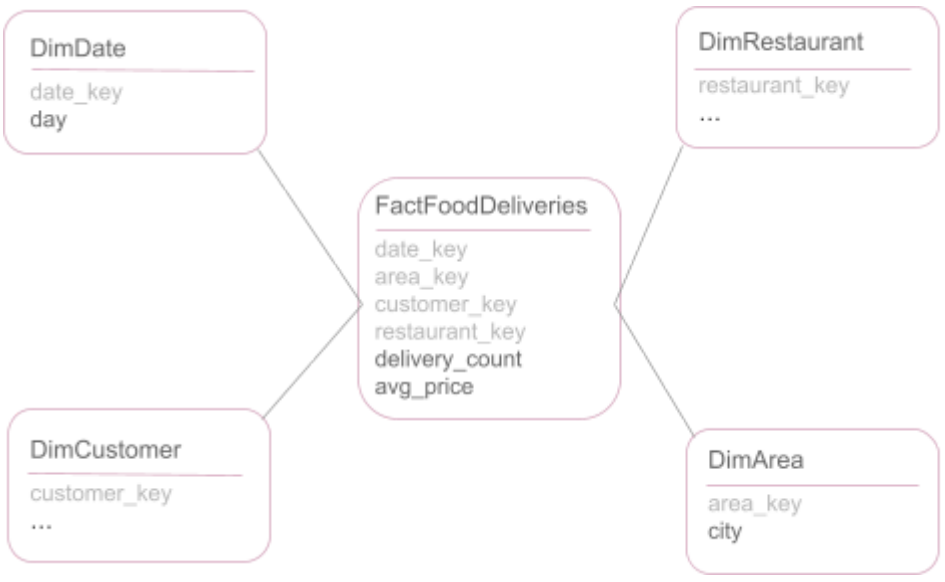
Assignment 5

1.1 Data Warehousing

A natural starting point for this schema would be the fact table. It needs to contain facts about the number of deliveries, average prices, and number of customers to be able to answer the given questions for the analysis.

How this schema is designed is highly dependent on what data is found useful, to answer the four example questions a simple star-schema would suffice; as it gives all the necessary information for them. On the other hand if more questions are to be answered an upgrade to a snowflake-schema with sub-dimensions would be more useful.

A. Assuming these four examples are the fundamentally needed ones, the star-schema could be as following:



These dimensions will allow us to filter on them to achieve the answer we seek from the example questions. More information on the dimensions could be useful depending on other queries, though for this particular case it will not be.

B. For the given concept hierarchy and dimensions only cuboid number two would be able to process the query. This is because it is the only cuboid containing the dimensions we need to search in, in this case item and location, filtering on time. Cuboid one and two would not be able to do this as they are both missing the necessary dimensions for our query.

2.1 Association Rules

TID	Transaction
1	A, B, C
2	A, B
3	A, D, E
4	D, E
5	E, C
6	A, E, D

1-Item	Support
A	4/6
B	3/6
C	3/6
D	3/6
E	3/6

Every item is frequent

2-Item	Support
AB	2/6
AC	2/6
AD	2/6
AE	2/6
BC	2/6
BD	0/6
BE	0/6
CD	0/6
CE	1/6
DE	3/6

In this case, AB, AC, AD, AE, BC and DE are frequent

3-Item	Support
ABC	2/6
ADE	2/6

In this case ABC and ADE is frequent

Rule	Confidence
A → BC	2/4 = 50%
B → AC	2/2 = 100%
C → AB	2/3 = 66%
A → DE	2/4 = 50%
D → AE	2/3 = 66%
E → AD	2/4 = 50%
A B → C	2/2 = 100%
AC → B	2/2 = 100%
BC → A	2/2 = 100%
AD → E	2/2 = 100%
AE → D	2/2 = 100%
DE → A	2/3 = 66%

Thus the valid associations rules become:

- B → AC
- C → AB
- D → AE
- AB → C
- AC → B
- BC → A
- AD → E
- AE → D
- DE → A

3.1 Decision Trees

1. Tables for calculating the GINI index for various attributes:

Age:

	Young	Middle	Old
Yes	4	5	3
No	3	1	4

Gini Index:

$$GINI = \frac{7}{20}\left(1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2\right) + \frac{6}{20}\left(1 - \left(\frac{5}{6}\right)^2 - \left(\frac{1}{6}\right)^2 + \frac{7}{20}\left(1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.426\right.\right.$$

Income:

	Low	Medium	High
Yes	4	6	2
No	2	3	3

$$GINI = \frac{6}{20}\left(1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2\right) + \frac{9}{20}\left(1 - \left(\frac{6}{9}\right)^2 - \left(\frac{3}{9}\right)^2 + \frac{5}{20}\left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.453\right.\right.$$

Married:

	Yes	No
Yes	9	5
No	3	3

$$GINI = \frac{12}{20}\left(1 - \left(\frac{9}{12}\right)^2 - \left(\frac{3}{12}\right)^2\right) + \frac{8}{20}\left(1 - \left(\frac{5}{8}\right)^2 - \left(\frac{3}{8}\right)^2 = 0.413\right.$$

TopGPA:

	Yes	No
Yes	8	4
No	4	4

$$GINI = \frac{12}{20}\left(1 - \left(\frac{8}{12}\right)^2 - \left(\frac{4}{12}\right)^2\right) + \frac{8}{20}\left(1 - \left(\frac{4}{8}\right)^2 - \left(\frac{4}{8}\right)^2 = 0.467\right.$$

Creditworth:

	Pass	High
Yes	6	6
No	4	4

$$GINI = \frac{10}{20}\left(1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2\right) + \frac{10}{20}\left(1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 = 0.480\right.$$

Whole set (yes/no):

Yes	12
No	8

$$GINI = 1 - \left(\left(\frac{12}{20}\right)^2 - \left(\frac{8}{20}\right)^2\right) = 0.480$$

2. Without using grouping during splitting, the marital status has the lowest Gini index indicating it would be the best splitting point.

3.

- Young, Medium Income, Top GPA, High credit: Yes
- Young, Low income, Low GPA, Pass Credit: No

4.1 Data Types

A. Continuous, Quantitative, Ratio

B. Continuous, Quantitative, Ratio

C. Discrete. Qualitative, Nominal (Arguably arbitrary number)

D. Discrete, Qualitative, Ordinal (Quantitative, Interval depending on interpretation)

E. Discrete, Quantitative, Ratio

F. Continuous, Quantitative, Interval

G. Discrete, Quantitative, Ratio

H. Continuous, Quantitative, Interval

I. Discrete, Quantitative, Ratio

J. Qualitative, Nominal (Can flag as Binary?)

K. Level of education could be represented multiple ways, Discrete, Qualitative, Ordinal is one option, it could also be represented as Qualitative, Nominal.

L. Discrete, Quantitative, Ratio

5.1 Noise and Outliers

A. Outliers can be interesting in for example medical sciences to look for discrepancies from normal values.

B. In theory you could remove all values equating to noise, but their frequency might make this undesirable or irrelevant. Practically it would not be possible to completely remove noise from a dataset.

C. Noise can skew and distort datasets, obscuring valuable data by removing accuracy.

D. The biggest tools against outliers and noise are the data cleaning and preprocessing steps.

E. As mentioned in A, both noise and outliers can be valuable. In addition to this they can help create more robust datasets conforming to the real world.

F. Yes, it would be possible for noise to be an outlier. Though not all noise will be an outlier.

G. Noise objects are not necessarily outliers. Often noise is not outliers.

H. No, outliers are usually not noise, but rather rare occurrences in the dataset.

I. Yes, enough noise can make the dataset skew either way.

6.1 Similarity Measures

1. ([1, 2], [3, 12])

Cosine: $sim = \frac{1 \cdot 3 + 2 \cdot 12}{\sqrt{1^2 + 2^2} \cdot \sqrt{3^2 + 12^2}} = 0.976$

Euclidean: $dist = \sqrt{(3 - 1)^2 + (12 - 2)^2} = 10.198$

Jaccard: $jacc = \frac{NULL}{1, 2, 3, 12} = 0$

2. ([1, –9, 23], [--1, –1, 12])

Cosine: $sim = \frac{1 \cdot (-1) + (-9) \cdot (-1) + 23 \cdot 12}{\sqrt{1^2 + (-9)^2 + 23^2} \cdot \sqrt{(-1)^2 + (-1)^2 + 12^2}} = 0.951$

Euclidean: $dist = \sqrt{(-1-1)^2 + (-1+9)^2 + (12-23)^2} = 13.748$

Jaccard: $jacc = \frac{NULL}{1, -9, 23, -1, 12} = 0$

3. ([−23,21,2,4], [−1,−9,2,−12])

Cosine: ≈ - 0.440

Euclidean: ≈ 40.497

Jaccard: $jacc = \frac{2}{-23, 21, 2, 4, -1, -9, -12} = 14.29\%$

4. ([−0.2,−0.2, 0.2], [−0.3,−0.1,−0.9])

Cosine: ≈ - 0.303

Euclidean: ≈ 2.777

Jaccard: 0

5. ([1,2,1,8,9,12,1,2] ,[12,9,3,2,1,0,3,3])

Cosine: ≈ 0.241

Euclidean: ≈ 20.567

Jaccard: $jacc = \frac{1, 2, 9, 12}{0, 1, 2, 3.8.9.23} = 57.14\%$

6. ([0.1,0.2], [0.3,0.4])
Cosine: = 1
Euclidean: ≈ 0.283
Jaccard: 0

7. ([0.5,0.6,0.7], [0.8,0.9,1.0])
Cosine: ≈ 0.999
Euclidean: ≈ 0.519
Jaccard: 0

8. ([0.12,0.34,0.56,0.78], [0.91,0.23,0.45,0.67])
Cosine: ≈ 0.757
Euclidean: ≈ 0.813
Jaccard: 0

9. ([1,−97,23], [1,−97,23])
Cosine: 0
Euclidean: 0
Jaccard: 100%

10. ([1,0,1,1,0], [1,0,1,1,1])
Cosine: ≈ 0.866
Euclidean:
Jaccard: 80%

11. ([1,1,0,1,0,1], [1,1,1,0,0,1])
Cosine: = 0.750
Euclidean:
Jaccard: 50%