Data-Warehouse-Technologien
Exercise sheet 6

Assignment 1: Which types of data quality errors can be found in the relation Customer and Order?

CID	Last Name	First Name	Address	Town	Birthday	contract capable
555666	Maier	Thomas	First Avenue 12	New York	1983-10-10	true
123456	Muster	Max	Rue du Tour 1	Frankreich	1972-01-01	true
112233	Schulz	Maik	M.-Gorki-Str. 5	Machteburch	2000-12-03	true
445566	Thomas	Maier	Rue du Gare 11	Paris	NULL	true
123456	Schulz	Mike	Maxim-Gorki-Strasse 5		1985-08-08	true

Table 1: Relation Customer

OID	CID	Article	Amount	Delivered
125	555666	4123649700201	1	T
512	123456	4222451689005	Zwei	1
699	112233	40815487990	3	0
730	555566	4900043174599	6	Nein
938	123456	3900004433901	Eins	Ja

Table 2: Relation Order

Classify the found errors and describe how these problems can be avoided/identified?

Assignment 2: Explain the problem of duplicate detection? Which methods can be used to detect duplicates? Which difference exists to the differential snapshot problem?

Assignment 3: Explain the differences between the ETL and ELT process.

Assignment 4: Create the schema of the TPC-H Benchmark by using the given script. Import the provided TPC-H Data into the data warehouse using the sqlldr. Consider the additional hints!

Assignment 5: Find all customer names (C_NAME) in the created data warehouse with an edit-distance from 1 to customer 'Customer#000000105'.

Assignment 6: Find all commentary in the created data warehouse (L_Comment) of orders (LINEITEM) with a soundex of 'W153'.
How was the soundex created?