Modul 100: Daten analysieren und strukturieren

Datenstrukturen und Datenwerte analysieren

Betrachten wir die Goldpreisentwicklung:

Die Tabelle (Entitätsmenge) `GOLDPREIS` besteht aus den drei Attributen `#ID, Jahr, USD.`

Tabellenschreibweise:
`TABELLENNAME(#PRIMARY KEY, Attr 1,..., Attr n)`
Beispiel:
`GOLDPREIS(#ID,Jahr,USD)`

Die Entitätsmenge `GOLDPREIS` hat 19 Tupel (Entitäten).
Jede Datenzelle hat einen Wert, es gibt keine leeren Zellen. Es gibt keine nicht atomaren Datenwerte.

Es fällt auch auf, dass das Attribut `ID` die Aufzählung 1 bis 19 als Datenwerte speichert.
1 bis 19 bedeutet, dass

- jeder Datenwert des Attributs `ID` **eine ganze Zahl (`INTEGER`)** ist.
- jede Zahl genau **einmal** vorkommt. Jeder Datenwert von `ID` ist eindeutig oder es gibt **keine doppelten Datenwerte**.

`ID` ist ein geeigneter `PRIMARY KEY` der Tabelle `GOLDPREIS.`

Das Attribut `Jahr` beinhaltet Jahresangaben und das Attribut `Preis` den Höchstpreis einer Unze Gold in Dollar.
Beachten Sie, dass die Aufzählung der Datenwerte des Attributs `Jahr` korrekt ist.
Die Tabelle hat eine **gute Datenqualität** hinsichtlich Datenstruktur und ist inhaltlich auch verifizierbar.

Beispielsweise:
http://www.fazfinance.net/Rohstoffe/Gold/XC0009655157/Metals/Wertpapier.html

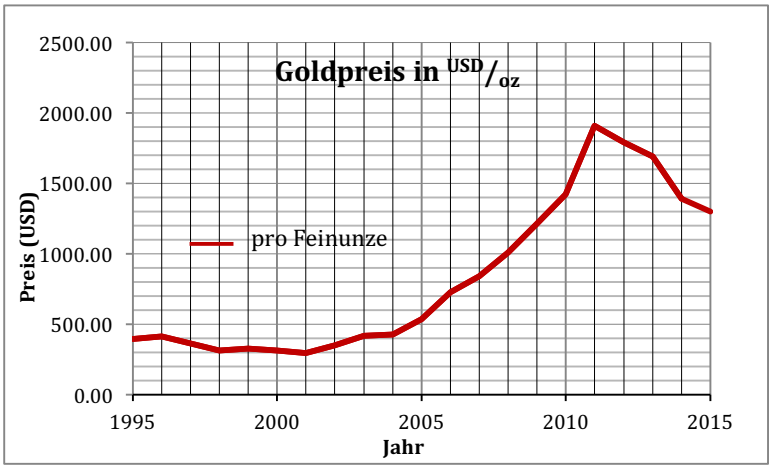
Jahreshöchstkurs einer Feinunze (oz) Gold (1oz = 31.1g) in US Dollar von 1995 bis 2015.

GOLDPREIS

#ID	Jahr	USD
1	1995	395.55
2	1996	414.80
3	1997	362.15
4	1998	313.15
5	1999	325.50
6	2000	312.70
7	2001	293.25
8	2002	349.30
9	2003	417.25
10	2004	427.25
11	2005	536.50
12	2006	725.00
13	2007	841.10
14	2008	1011.25
15	2009	1212.50
16	2010	1421.00
17	2011	1908.79
18	2012	1793.83
19	2013	1693.75
20	2014	1390.47
21	2015	1302.30

Quelle: finanzen.net

Veranschaulichen wir die Daten in einem Liniendiagramm:



Redundanzen

Betrachten wir die Tabelle `MYFRIENDS` (meine Freunde), die auf einem Datenbankserver abgelegt ist.

#ID	Name	Vorname	Strasse	Ort	email
1	Turner	Tina	Goldstreet	8001 Zuerich	ttt@pop.com
2	Maurer	Ralph	Greenstreet	8000 Zurich	ralph.maurer@gibb.ch
3	Frei	Beat	Bluestreet	8002 Zurich	beat.frei@gibb.ch
4	Maurer	Ralph	Greenroad	3005 Bern	ralph.maurer@gibb.ch
5	Bernanke	Ben	Blackstreet	New York	bb@fed.us.gov
6	Meier	Claudia	Brownhouse	3005 Berne	mc@meier.ch
7	Meier	Claudia	Brownhouse	3001 Bern	mc@meier.ch

Alle Attribute haben eindeutige Namen.

Tabellenschreibweise:
`MYFRIENDS(#ID PRIMARY KEY, Name, Vorname, Strasse, Ort, email)`

Die Datenstruktur der Tabelle scheint „vorläufig“ in Ordnung zu sein. Aber wie sieht es mit den Datenwerten aus?

Wir stellen schnell fest, dass die Datensätze mit `ID = 2` und `ID = 4` sowie `ID = 6` und `ID = 7` die gleichen Datenwerte in den Attributen `Vorname, Name` und `email` haben. Falls nur die Namen und Vornamen übereinstimmen würden, könnte man von unterschiedlichen Personen mit gleichen Vornamen und Namen ausgehen. Unterschiedliche Datenobjekte mit gleichen Namen. Da aber auch die Emailadressen gleich sind, handelt es sich um **Redundanzen** und um eine **schlechte Datenqualität** in der Tabelle `MYFRIENDS`.

Redundanzen führen in Tabellen zu Problemen
Das Problem mit Redundanzen lässt sich einfach zusammenfassen. Sie verursachen folgende Situationen:

- Aufgrund mehrfach vorhandener Datensätze mit gleichen Datenwerten, weiss man als Datennutzer nicht, welcher Datensatz der Aktuellste ist.
- Möchte ich dem Datensatz eine weitere Information hinzufügen, weiss ich nicht, bei welchem Datensatz ich das tun muss.
- Will ich einen redundanten Datensatz löschen, weiss ich nicht, ob ich den richtigen oder gar verschiedene Datensätze zum Löschen selektiert habe.

Angenommen ich will Herrn Ralph Maurer einen Brief schreiben, welche Adresse muss ich nun verwenden? Die Adressangaben beider Datensätze sind verschieden.
Bis jetzt betrachteten wir Redundanzen aus der Perspektive aller Attribute eines Datensatzes. Aber wie sieht es mit einem einzelnen Attribut aus?

Untersuchen wir das Attribut `Ort` und seine Datenwerte:
Allein die Tatsache, dass die Datenwerte Zuerich und Zurich sowie Bern und Berne zweimal unterschiedlich geschrieben sind, lässt auf eine ähnliche Problematik wie oben hindeuten.

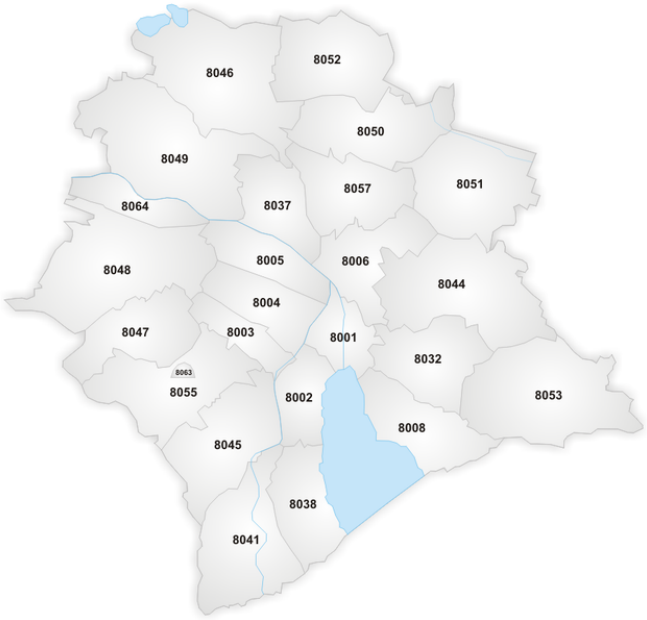
Wird mit einem Datenwert das gleiche Datenobjekt beschrieben, wie in unserem Fall Bern oder Zürich, sollten die Ortsbezeichnungen gleich sein oder man läuft Gefahr

- nur bei Bern statt auch Berne die PLZ zu aktualisieren oder
- keine Treffer bei einer Suche nach Freunden in Zürich zu erhalten (mit „ü“ geschrieben).

Konsistenzen

Das Attribut `Ort` enthält Datenwerte von geringer Datenqualität, aber selbst dann, wenn die Datenwerte bereinigt sind, können wir folgendes feststellen:

Unbereinigt:	Bereinigt:
8001 Zuerich	8001 Zürich
8000 Zurich (**ungültig**)	8001 Zürich
8002 Zurich	8002 Zürich
3005 Bern	3005 Bern
New York	New York US.
3005 Berne	3005 Bern
3001 Bern	3001 Bern



PLZ der Stadt Zürich

Korrekt wäre 8001 Zürich und 8002 Zürich als zwei unterschiedliche Datenobjekte zu führen, da es ja auch unterschiedliche Datenobjekte der realen Welt sind.
Das gleiche gilt für die Stadt Bern. Es gibt sowohl 3001 Bern und 3005 Bern.

Der Datenbankdesigner muss sich entscheiden, ob er die Ortsbezeichnungen in Deutsch oder/und in Französisch führen möchte. Berne ist die französische Bezeichnung für Bern. Viele Datenbanksysteme führen Schweizer Ortschaften in allen Schweizer Landessprachen.

Betrachten wir nun das Attribut `Ort` bereinigt in der Tabelle `MYFRIENDS`:
In der vorliegenden Datenstruktur der Tabelle `MYFRIENDS` führt das Attribut Ort zu vielen redundanten Datenwerten.

Sogar, wenn die redundanten Datensätze mit `ID = 2` und `ID = 7 gelöscht` sind, beinhaltet die Tabelle nach wie vor zweimal die Datenwerte 3005 Bern:

#ID	Name	Vorname	Strasse	Ort	email
1	Turner	Tina	Goldstreet	8001 Zürich	ttt@pop.com
3	Frei	Beat	Bluestreet	8002 Zürich	beat.frei@gibb.ch
4	Maurer	Ralph	Greenroad	3005 Bern	ralph.maurer@gibb.ch
5	Bernanke	Ben	Blackstreet	New York	bb@fed.us.gov
6	Meier	Claudia	Brownhouse	3005 Bern	mc@meier.ch

Beispiel: Denken wir weiter, habe ich 48 Freunde in 3005 Bern, bedeutet dies, dass die Tabelle `MYFRIENDS` 48 mal den Datenwert 3005 Bern führt.

Um Redundanzen in den Datenwerten von Attributen zu vermeiden, lagert man diese in eine separate Tabelle `ORT` aus und verbindet die Tabellen über einen **Fremdschlüssel**, der auf einen **Primärschlüssel in der Primärtabelle referenziert**.

Beispiel:
Fremdschlüsseltabelle:
`MYFRIENDS(#ID PRIMARY KEY, Name, Vorname, Strasse, email, #FKORTSID FOREIGN KEY)`

#ID	Name	Vorname	Strasse	email	#FKORTSID
1	Turner	Tina	Goldstreet	ttt@pop.com	1
3	Frei	Beat	Bluestreet	beat.frei@gibb.ch	2
4	Maurer	Ralph	Greenroad	ralph.maurer@gibb.ch	3
5	Bernanke	Ben	Blackstreet	bb@fed.us.gov	4
6	Meier	Claudia	Brownhouse	mc@meier.ch	3

Primärschlüsseltabelle:
`ORT(#ORTSID PRIMARY KEY, PLZ, Ort)`

#ORTSID	PLZ	Ort
1	8001	Zürich
2	8002	Zürich
3	3005	Bern
4		New York
5	3001	Bern

Die Unterteilung der Tabelle `MYFRIENDS` in zwei Tabellen `MYFRIENDS` und `ORT` schafft in diesem Fall **Konsistenz**.

Hierzu müssen folgende Bedingungen erfüllt sein:

- Jeder Datenwert des Fremdschlüssels `FKORTSID` ist in der Primärschlüsseltabelle `ORT` auch als Datenwert des Primärschlüssels vorhanden.
- Der Fremdschlüssel muss vom gleichen Datentyp wie der Primärschlüssel sein. In anderen Worten darf der Primärschlüssel nur Zahlen (`INTEGER`) und der Fremdschlüssel auch nur Integer-Zahlen beinhalten. Es ist nicht erlaubt, dass der Fremdschlüssel aus Buchstaben und der Primärschlüssel aus Zahlen besteht.

Definition I: Konsistenz
Konsistenz (consistency, semantic integrity) ist die Freiheit von Widersprüchen innerhalb einer Datenbank. Diese Widerspruchsfreiheit ist dann gegeben, wenn alle Redundanzen beseitigt sind.

Definition II: Strukturierte Daten
Ein Datenbestand heisst strukturiert, wenn systematische Untergliederungen und Verknüpfungen möglich sind. Nur strukturierte Daten erlauben effizientes Suchen und Bearbeiten von grossen Datenmengen.

Definition III: Konsistenz
Konsistenz meint die Übereinstimmung von an verschiedenen Stellen gespeicherten Daten oder des Bezugs zwischen Daten in der Datenablage. Als Inkonsistenz werden einander widersprechende Daten bezeichnet (inkonsistent = widersprüchlich). Inkonsistenzen führen zu falschen Ergebnissen bei der Arbeit mit Datenbanken.

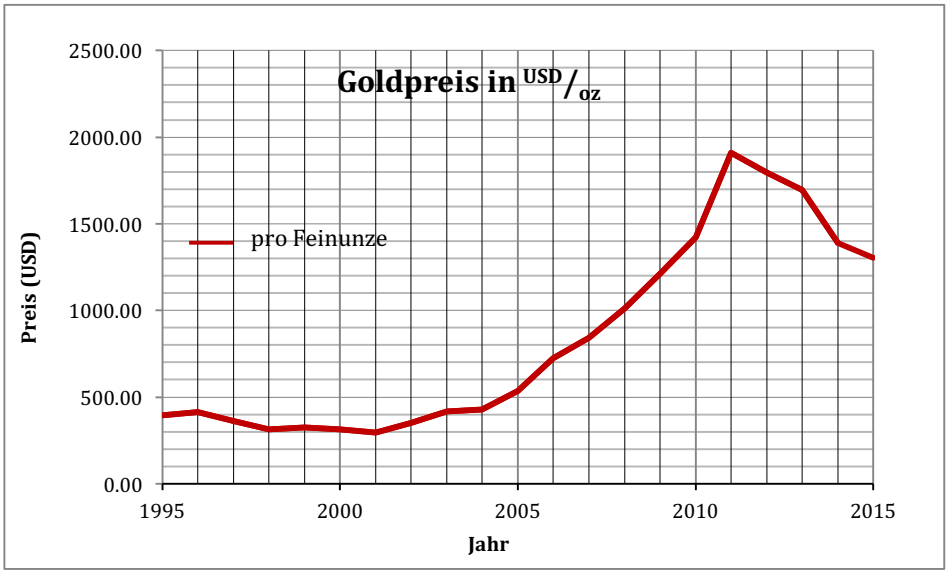
Repetitorium AB100-03

Aufgabe 1

Werten Sie das Kurvendiagramm Preis [1oz/USD] aus:

Nennen Sie mind. 5 Erkenntnisse / Informationen, die Sie aus den Datenwerten ableiten können.

Tipp: Betrachten Sie die Datenreihe (Kurve) und versuchen Sie in Abhängigkeit zu den Datenwerten der Y- und X-Achse Schlüsse zu ziehen.



Aufgabe 2

Grundkenntnisse: Datencharakterisieren

Begriffe:

A	Fremdschlüssel	**B**	Referenz
C	Redundanz	**D**	Konsistenz
E	Dokumente	**F**	Datenqualität
G	Strukturierte Daten	**H**	Datentyp
I	Inkonsistenz	**J**	INTEGER

Welcher der oben aufgeführten Begriffe passt jeweils am besten zu den unten aufgeführten Beschreibungen? Mehrfachnennungen sind möglich (ein Begriff passt zu Beschreibungen).

Beschreibungen:

Verbindung zwischen zwei Tabellen.
A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐ I ☐ J ☐

Kommt irgendwo als Primärschlüssel vor.
A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐ I ☐ J ☐

Integer (Zahl) ist ein ...
A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐ I ☐ J ☐

Bezeichnung des Datentyps ganzer Zahlen.
A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐ I ☐ J ☐

Gleiche Datenwerte in einem Attribut nennt man auch ...
A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐ I ☐ J ☐

Widerspruch in der Referenz zwischen zwei über Primär- und Fremdschlüssel referenzierte Tabellen.
A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐ I ☐ J ☐

Daten, die sich referenzieren lassen.
A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐ I ☐ J ☐

Widerspruchsfreie Datenbestände sind ...
A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐ I ☐ J ☐

Unstrukturierte Daten entsprechen ...
A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐ I ☐ J ☐

Der Datentyp einer ganzen Zahl ...
A ☐ B ☐ C ☐ D ☐ E ☐ F ☐ G ☐ H ☐ I ☐ J ☐

Aufgabe 3

a) Analysieren Sie die referenzierten Tabellen `PERSON` und `ORT` auf Datenstruktur und Datenqualität.

b) Gibt es Inkonsistenzen?
Inkonsistenzen sind Fehler, die durch falsche Referenzen entstehen können.

PERSON

PID	Name	Vorname	FK_ORTSID
1	Pignone	Clara	4
2	Fischer	Fabio	5
3	Burri	Sandro	11
4	Zubler	Sandrine	
5	Frutschi	Lea	3
6	Teuffer	Tobias	6
7	Schaedler	Thomas	9
8	Pigic	Nair	10
9	Frutschi Lea		+1
10	Heynen	Frieda	2

ORT

ORTSID	PLZ	Ortsname
1	1700	Freiburg
2	3280	Murten
3	3005	Bern
4	8001	Zürich
5	3110	Münsingen
5	3074	Muri
7	3285	Galmiz
8	3018	Bümpliz
9	3005	Wabern
10	3280	Murten

Aufgabe 4

Finden Sie in den folgenden Tabellen die verschiedenen Datenfehler?

a) Markieren Sie die Fehler in der Tabelle.

b) Erklären Sie jeden Fehler kurz.

tblArtikel : Tabelle

AID	MID	AName	Gewicht	AnzahlAnLager	OID
1	1	Schokohase	200	10	1
2	4	Weihnachtshase	1	2	7
3	7	Langohrhase	2	0	9
4	8	Allhase	11111111111		2
5	5	Zahnloshase	420024	5	4
6	2	Obdachhase	0	6	
7	9	Flughase	66	7	2

tblVerkauf : Tabelle

AID	KID	Datum	Anzahl	Bemerkungen
1	6	02.01.2009	14	jede Woche zwei Stück
2	3	01.01.2009	5	aber diesmal pünktlich!
2	2	04.01.2009	3	lieber an Ostern
3	3	03.01.2009	0	nicht schon wieder
4	4	06.01.2009	24	mit Dreikönigskuchen
5	4	07.01.2009	20	ohne MWSt
8	1	05.01.2009	3	per Nachname

tblKunden : Tabelle

KID	KName	KVorname	OID
1	Bonzo	Bonzius	1
2	Zahlung	Schwach	1
3	Kredito	Immer	9
4	Solvenzo	No_Go	
5	Zinso	Wucher	2

tblMaterial : Tabell

MID	Material
1	Schokolade
1	Marzipan
3	Stoff
4	Filz
5	Karton
6	Uran239

tblOrt : Tabelle

OID	OName
1	Runaway
2	Underbridge
3	Paradise
4	Anderswo
5	Chaos
6	Ghetto
7	Submarine
8	Heaven

Fehlerbeschreibung: