Appendix B
Examples of Data Quality Issues

In Section 3.4, we briefly discussed a taxonomy of data quality problems. This appendix provides a list of concrete examples for all categories of problems from the taxonomy. We start with single-source problems followed by multi-source problems.

B.1 Single-Source Problems

B.1.1 Missing Data

Missing Value

- Missing time/interval and/or missing value
  (Date: NULL, items-sold: 20)
- Dummy entry
  (Date: 1970-01-01); (duration: -999)

Missing Tuple

- Missing time/interval + values
  (The whole tuple is missing)

B.1.2 Duplicate Data

Unique Value Violation

- Exact same time/interval although time/interval is defined as unique value
  (Holidays: 2012-04-09; 2012-04-09)

W. Aigner et al., *Visualization of Time-Oriented Data*, Human–Computer Interaction Series, https://doi.org/10.1007/978-1-4471-7527-8

Exact Duplicates

- Same time/interval and same values
  (Date: 2012-03-29, items-sold: 20 is in table twice)

Inconsistent Duplicates

- Same real entity with different times/intervals or values
  (patient: A, admission: 2012-03-29 8:00) vs.
  (patient: A, admission: 2012-03-29 8:30)
- Same real entity of time/interval (values) with different granularities (rounding)
  (Time: 11:00 vs. 11:03); (Weight: 34,67 vs. 35)

B.1.3 Implausible Data

Implausible Range

- Very early date/time in the future
  (Date: 1899-03-22); (date: 2099-03-22); (date: 1999-03-22, duration: 100y)

Unexpected Low/High Values

- Deviations from daily/weekly... profile or implausible values
  (Average sales on Monday: 50) vs. (this Monday: 500)
- Changes of subsequent values implausible
  (Last month: 4000 income) vs. (this month: 80000 income)
- Too long/short intervals between start–start/end–end
  Below one second at the cash desk
- Too long/short intervals between start–end/end–start
  Off-time between two shifts less than 8h
- Too long/short overall timespan (first to last entry)
  Continuous working for more than 12 hours
- Same value for too many succeeding records
  17 customers in every interval of the day

B.1.4 Outdated Data

Outdated Temporal Data

- Only old versions available
  Sales values from last year
- New version replaced by the old version
  Project plan tasks overwritten by prior version

B.1.5 Wrong Data

Wrong Data Type

- No time/interval
  *Date: AAA; duration: **

Wrong Data Format

- Wrong date/time/datetime/duration format
  (Date: YYYY-MM-DD) vs. (date: YY-MM-DD); (duration: 7.7h) vs. (duration: 7h42')
- Times outside raster (e.g., for denoting end of the day)
  1-hour-raster but time is 23:59:00 for the end of the last interval

Misfielded Values

- Time in date field, date in time field/duration field
  (Time in date field: 14-03, date in timefield: 12:03:08)
- Values attached to the wrong/adjacent time/interval
  GPS data shows sprints followed by slow runs although the velocity was constant

Embedded Values

- Date+time in date field, time zone in time field/duration field
  (Time: 22:30) vs. (time: 22:30 CET)

Coded Wrongly or not Conform to Real Entity

- Wrong time zone
  UTC data in stead of local time
- Valid time/interval but not conform to the real entity
  (Admission: 2012-03-04) vs. (real admission: 2012-03-05)

Domain Violation (Outside Domain Range)

- Outliers in % of concurrent values (attention with small values) for a given point in time/interval
  On average (median) 30 customers in a shop in a given hour – in a 10' interval within that hour, a value of 200 is present
- Uneven or overlapping intervals
  Turnover data for 8:00–9:00, 9:00–11:00, 11:00–12:00
- Minimum/Maximum violation for given time/interval/type of day
  Sales at night even though no employees were present
- Sum of subintervals impossible
  Seeing the doctor + working hours longer than regular working hours
- Start, end, or duration do not form a valid interval
  (End ≤ start); (duration ≤ 0)
- Circularity in a self-relationship
  Interval A ⊂ interval B, interval B ⊂ interval A, A ≠ B

Incorrect Derived Values

- Error in computing duration
  Error computing sum of employees present within two intervals: (interval: 8:00–8:30, employees: 3), (interval: 8:30–9:00, employees: 3) → (interval: 8:00–9:00, employees: 6); no proper dealing with summer time-change; computing the number of work hours per day without deducting the breaks

B.1.6 Ambiguous Data

Abbreviations or Imprecise Unusual Coding

- Ambiguous time/interval/duration due to short format
  (Date: 06-03-05) vs. (date: 06-05-03); 5' interval encoded as '9:00': (interval: 8:55–9:00) vs. (interval: 9:00–09:05); average handling time per given interval: 3' – not clear: (average of completed interactions) vs. (average of started interactions) within this interval

- Extra symbols for time properties
  *+ or * or 28:00 for next day*

B.2 Multi-Source Problems

B.2.1 Heterogeneous Syntaxes

Different Data Formats/Synonyms

- Different date/duration formats
  (Date: YYYY-MM-DD) vs. (date: DD-MM-YYYY); (Date: 03-05 (March 5)) vs. (date: 03-05 (May 3))

Different Table Structure

- Time separated from date vs. date+time or start+duration in one column
  (Table A: start-date, start-time) vs. (table B: start-timestamp)

B.2.2 Heterogeneous Semantics

Heterogeneity of Scales (Measure Units/Aggregation)

- Different granularities; different interval length
  (Table A: whole hours only) vs. (table B: minutes)

Information Relates to Different Times/Intervals

- Different times/intervals
  (Table A: current sales as of yesterday) vs. (table B: sales as of last week)

B.2.3 References

Referential Integrity Violation/Dangling Data

- No reference to a given time/interval in another source
  (Table A: sales per day), (table B: sales assistants per day), problem: table B

does not contain a valid reference to a given day from table A or table A does not contain any referencing time

Incorrect Reference

- Reference exists in other sources but does not conform to real entity
  Sales of one day (table A) are assigned to certain sales assistants (from table B) because they reference the same day, however, in reality, a different crew was working on that day

B.3 Summary

With the help of the above examples, one can systematically check whether data quality problems exist. If some problems are overlooked, it is often the case that problems will pop up anyway when one tries to visualize incorrect data. This can often be recognized by artifacts in the visualization or algorithms failing to execute properly. Yet, verifying data correctness before the actual visualization is always preferable to stumble upon data problems during the data exploration.