Data complete

Various authors

2023-08-22

Table of contents

Introduction

This book will discuss electronic representations of data. Though this is intended as a practical guide for actuaries working with insurance and risk management data, the principles and techniques may be more general than that.

1 Data types

1.1 Basic data types

How computers represent floating point numbers.

Data type conversion.

1.2 Matrices and tensors

1.3 Structured and unstructured data

1.4 Structure/record/user-defined types

Not quite a thing anymore. More historical/background than practical.

1.5 List-like objects

Heterogeneous, recursive data, JSON/XML

1.6 Document-like objects

key-value pairs

2 Data frames

2.1 Normalization/tidy data

2.2 Base R

```r
df_example <- data.frame(
  col1 = 1:2,
  col2 = 3:4
)
```

2.3 Python

```python
import pandas as pd

df_example = pd.DataFrame(
  {'col1': [1, 2], 'col2': [3, 4]}
)
```

2.4 tidyverse

```r
library(tibble)

tbl_example <- tibble(
  col1 = 1:2,
  col2 = 3:4
)
```

2.5 CSV and pitfalls

2.6 Parquet

3 Queries

3.1 Columnar and row-wise subsets.

3.2 Logical and ordinal slicing.

3.3 Random sampling, stratified sampling.

4 Combining data frames

4.1 Inner join

4.2 Outer joins

Also cover semi-joins

4.3 Anti-joins

4.4 Rolling joins

4.5 Union/binding rows

4.6 Inserting single rows

5 Transforms

5.1 Mutations

5.2 Aggregation/summarization

5.3 Window functions

5.4 Pivoting/spreading/unstacking

6 EDA

6.1 Tabular/query

6.2 Visual

7 Missing data

7.1 Types

Missing-at-random, missing-not-at-random, etc.

7.2 How to address it

Talk to data collectors, ignore it, impute it,

8 Relational database management systems

8.1 Pour one out for Edgar Codd

8.2 Normalization

8.3 ACID, CRUD and all that

8.4 Superficial notes about what a DBA does

9 NoSQL

See Luc Perkins and Wilson (2018).

9.1 Graph databases

Neo4j

9.2 Document data

MongoDB

10 Big data

10.1 Definition, examples

Four V's, etc.

10.2 map/reduce

10.3 Spark

11 ETL, warehouses, etc.

Need a succinct title for this chapter. Basic idea is: how do multiple data systems interact? What is ETL, what the F$#% is a data lake and so on.

12 Conclusion

Data is hard.

References

Luc Perkins, Eric Redmond, and Jim Wilson. 2018. *Seven Databases in Seven Weeks 2nd Edition a Guide to Modern Databases & the NoSQL Movement.* The Pragmatic Programmers.