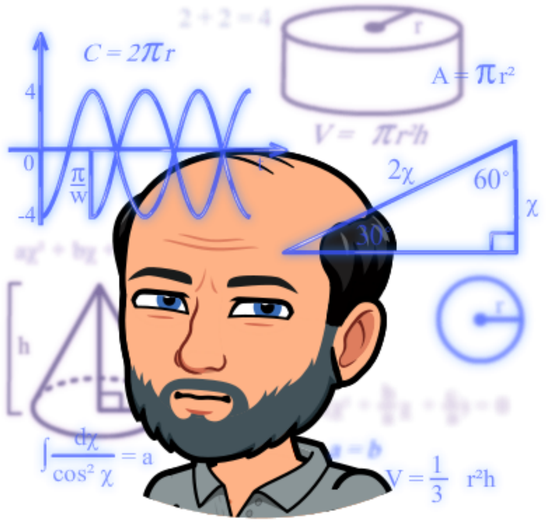
Data profiling done right from the start



Thomas Hütter

Data Innovation Summit 2023, Kista

Data profiling done right from the start

Thomas Hütter, Diplom-Betriebswirt

- Developer for ERP apps, SQL scripts and BI stuff
- Worked at consultancies, ISVs, end user companies
- 1995: SQL Server, 2001: Nav/BC, 2014: R, 2020: Power*
- Speaker at SQL / data / dev events around Europe

@DerFredo https://twitter.com/DerFredo

de.linkedin.com/in/derfredo

https://techhub.social/@DerFredo

sessionize MOST ACTIVE SPEAKER 2022

SWECON

Developer! DeveloperDeveloper UK Community Events

sqlbits

{ } NDC Conferences

DATA INNOVATION SUMMIT 2023

update

Agenda

- What does „data profiling“ mean?
- What does data profiling comprise?
- What does data profiling cost?
- Techniques & tools
- Some visualization examples
- Round-up, conclusions, resources

What does „data profiling“ mean?

- „The process of examining the data available from an existing information source (e.g. a database or a file) and collecting statistics or informative summaries about that data.“ [Wikipedia]

  … „Assess data quality, including whether the data conforms to particular standards or patterns“

  … „The result of the analysis is used to determine the suitability of the candidate source systems, usually giving the basis for an early go/no-go decision, and also to identify problems for later solution design.“

- „… is a technology for discovering and investigating data quality issues, such as duplication, lack of consistency, and lack of accuracy and completeness.“ [Gartner]

  … „enables the data steward to investigate the origin of data errors.“

- „Data profiling - having that first date with your data“ [Dr. Kirk Borne]

  … „the best path to ‘Knowing thy data’“, one fundamental principle of sound data science

Data profiling

- vs Data Mining:

  Data profiling gathers statistics, summaries on the structural consistency of your data, helping to determine the „fitness“ of the data to your purposes.

  Data mining aims to reveal „hidden“ patterns or relations within the data itself that can lead to non-obvious insights and thus use of your data.

- vs Data cleansing:

  A *data profiling* process may (should ?) result in the need for *data cleansing*, which is a separate process to deal with the problematic data items, i.e. correcting, consolidating, adding or deleting data to make it fit for your purposes.

Data profiling

- *Data Governance*: management concept implementing processes needed to ensure high data quality throughout the lifecycle of your data.
  Focuses on data consistency / integrity, usability, standards compliance etc.
  May be driven *internally*:
  - management wants reliable, correct data to base their decisions upon
  - stakeholders want optimal decisions to grow performance / profit

  or driven *externally*:
  - urge to comply with local, national, international laws and regulations
- *Data Quality* is considered high when the data is „fit for the intended purpose“, i.e. it correctly represents the real world circumstances.
  Thus your data has to be correct, consistent and complete etc.

Data profiling

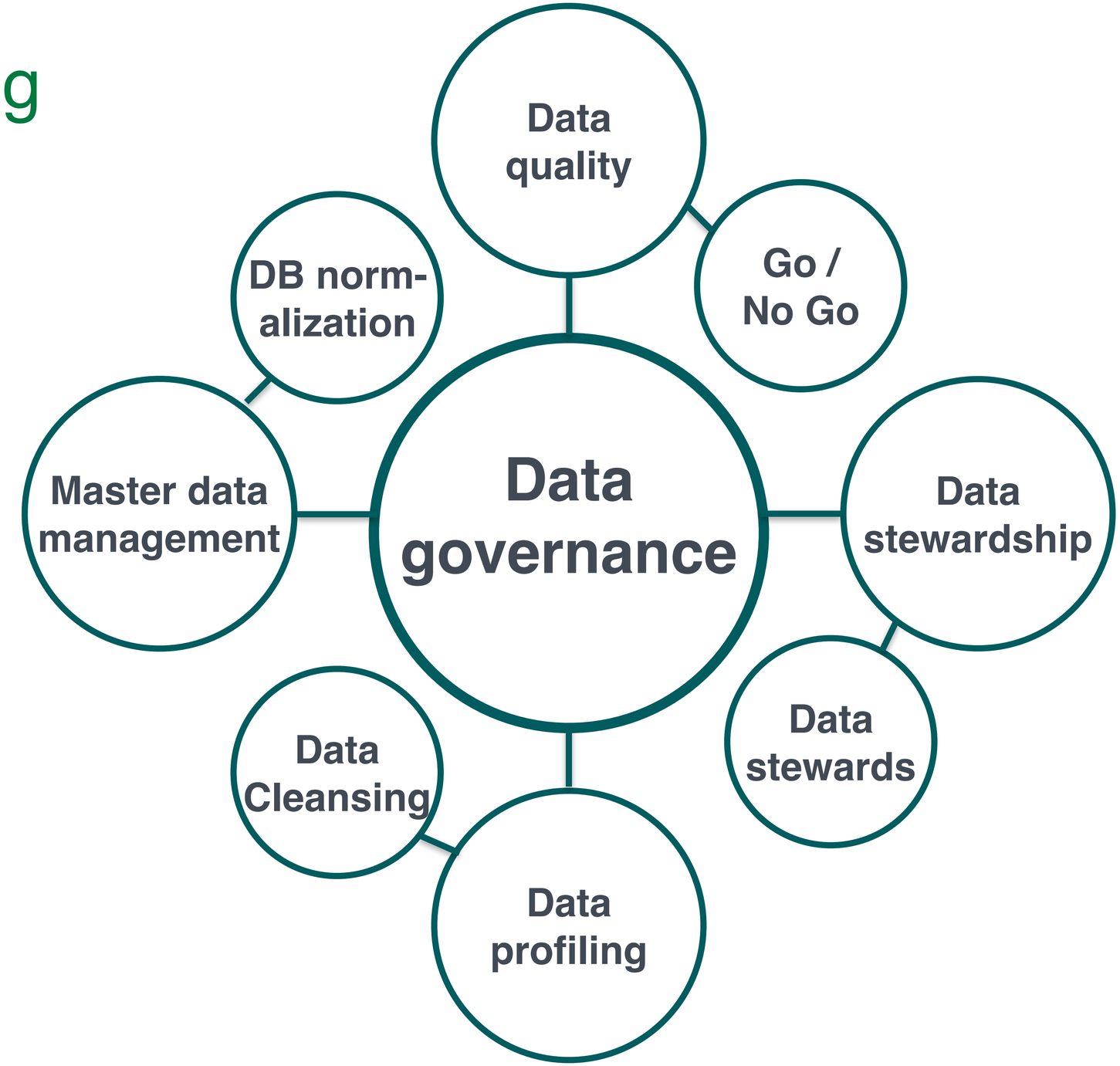
- *Master Data Management*:
  The discipline, taken out by data owners or data stewards, that ensures the overall accuracy of a company's master data assets.

  Avoid „garbage in - garbage out“ in your downstream processes.

- *Database normalization*:
  „Ted Codd's 12 commandments“, (actually 13) rules that define the requirements for a true relational database management system.

  Structuring your db according to so-called „normal forms“ 1 to 6.

Data profiling



What does data profiling comprise?

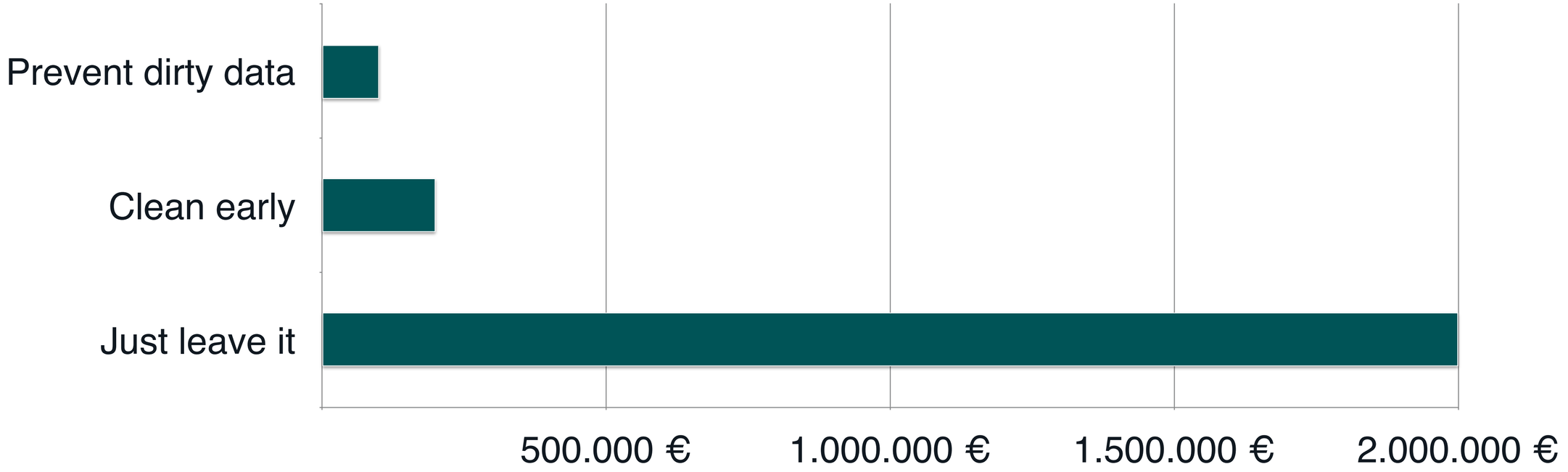
- Looking at tabular, relational data:
- Single column criteria:
  data type, cardinality / distribution, uniqueness / duplicates, empty or missing values, data range / outliers, key column …
- Multiple columns criteria:
  cardinality and uniqueness of combinations, combo restrictions, consistent use of columns, dependencies, key columns …
- Multiple sources criteria:
  schema match / mismatch, handling of duplicates, auxiliary data matching …

What does data profiling comprise?

- Data cleansing activities:
- Remove / consolidate duplicate records / values where they should be unique
- Add missing data, by inserting average (which?), zero or other values
- Determine handling of outliers: correct or drop
- Correct typos, inconsistent abbreviations or category names (manual data entry?)
- Care for / convert to up-to-date categorial data
- Think of dependent data when changing key fields
- Unify / consolidate data from multiple sources

What does data profiling cost?

- A long-standing theory amongst data experts is the „1-10-100 rule“.
- It takes $1 per record to prevent dirty data in first place, $10 to clean it early, but $100 in the long run, if you just leave it dirty like that.
  Assuming a 100,000 record database and 20% dirty records, this means:



What does data profiling cost?

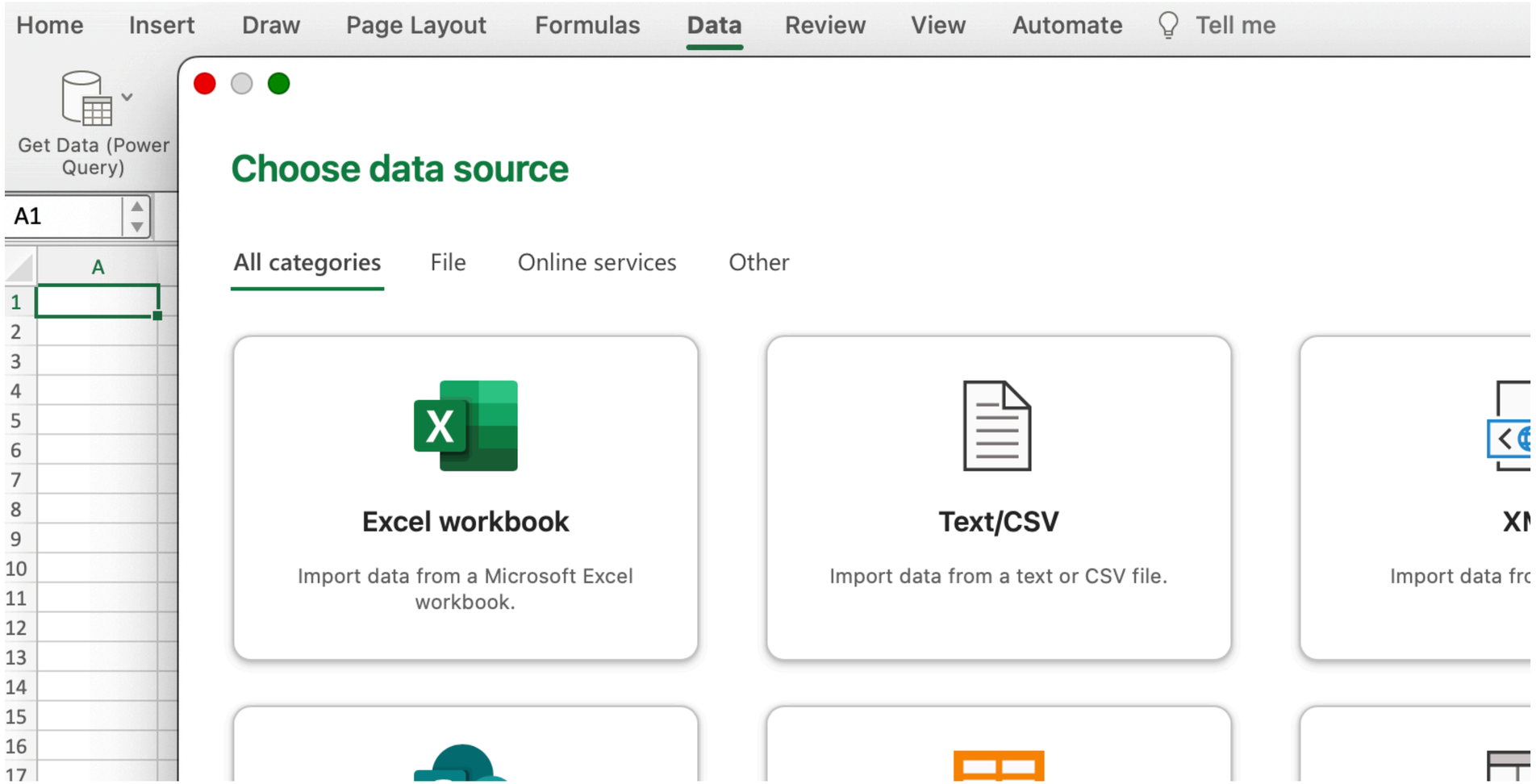
- According to a 2016 IBM estimate, the *yearly* cost of poor quality data for *US* companies alone adds up to 3.1 trillion (3,100,000,000,000) US-$. [HBR]
  Comparison: (IDC 2016) big data market worldwide: 136 billion US-$
  (Other sources: 5 to 15 million US-$ per year for an average organization)

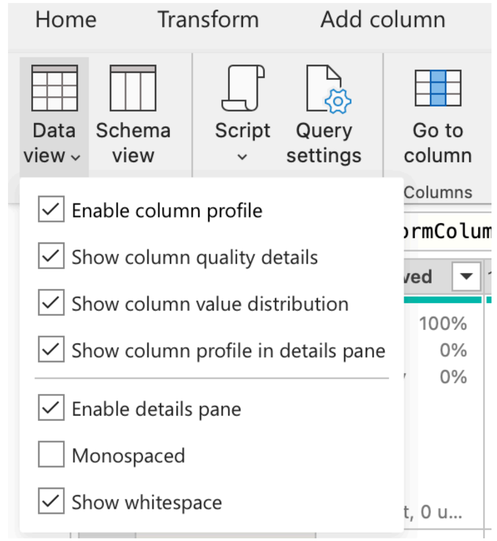
- How does dirty data cost your company money:
  - direct costs of handling the problems in the data
  - risc of fines for not obeying laws and regulations
  - lost sales opportunities, wasted procurement costs
  - possible interruption of your operation
  - loss of good reputation due to poor customer experience

Techniques and tools

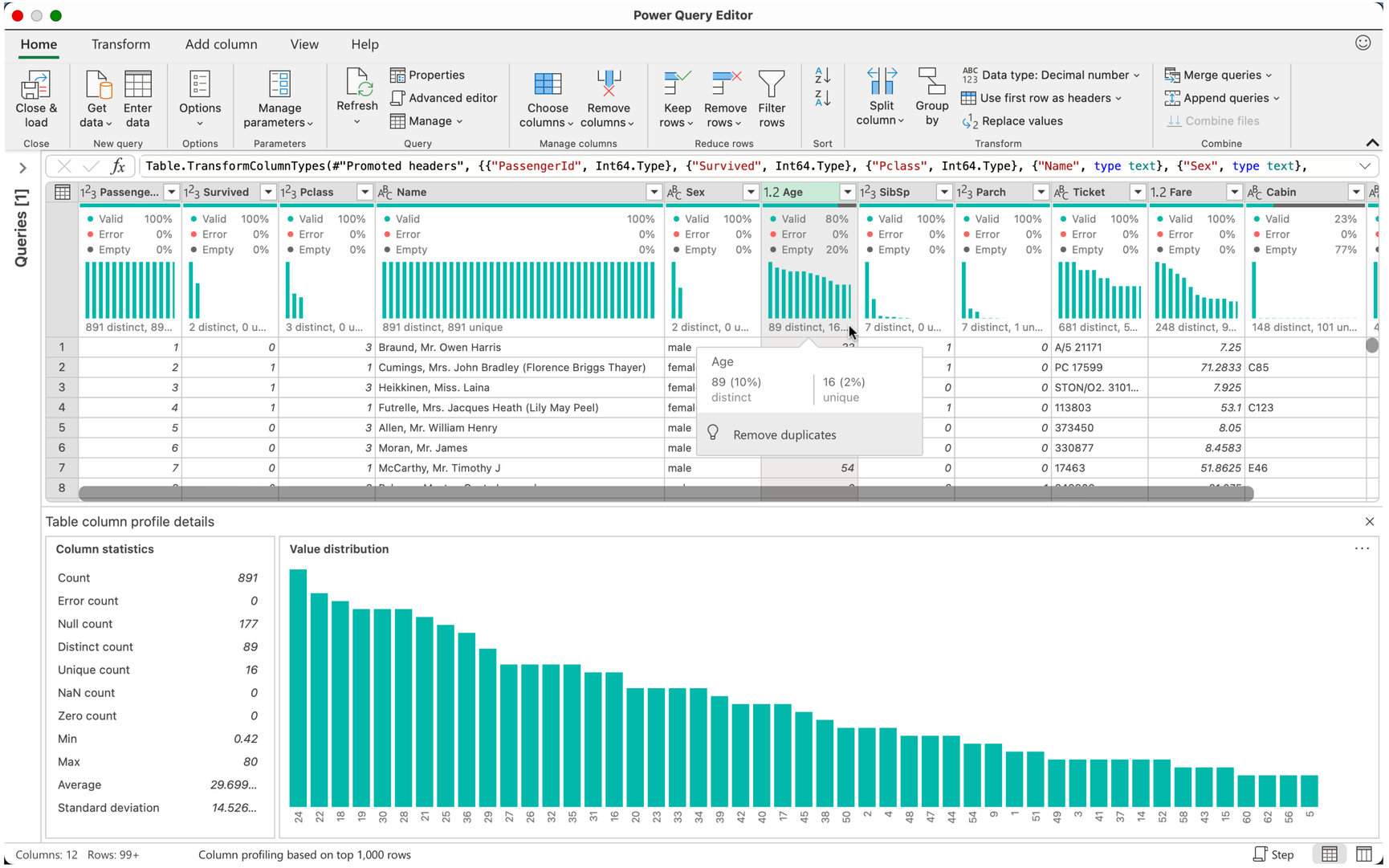
- There are countless tools available, commercial and open-source / free (some examples below; to list them all here can't be comprehensive) [dbmstools]
- Commercial:
  Atlan, Informatica data profiling, IBM Infosphere information analyzer, Melissa data profiler, Microsoft Excel, SSIS and Power BI, Oracle Enterprise Data Quality, SAS DataFlux, Talend Data fabric, Toad Data point …
- Open-source:
  Apache Griffin, DataCleaner, DataMatch Enterprise, Google OpenRefine, Idera SQL Data profiler, OpenDQ, Python, R/RStudio …

Techniques and tools: Excel / Power Query

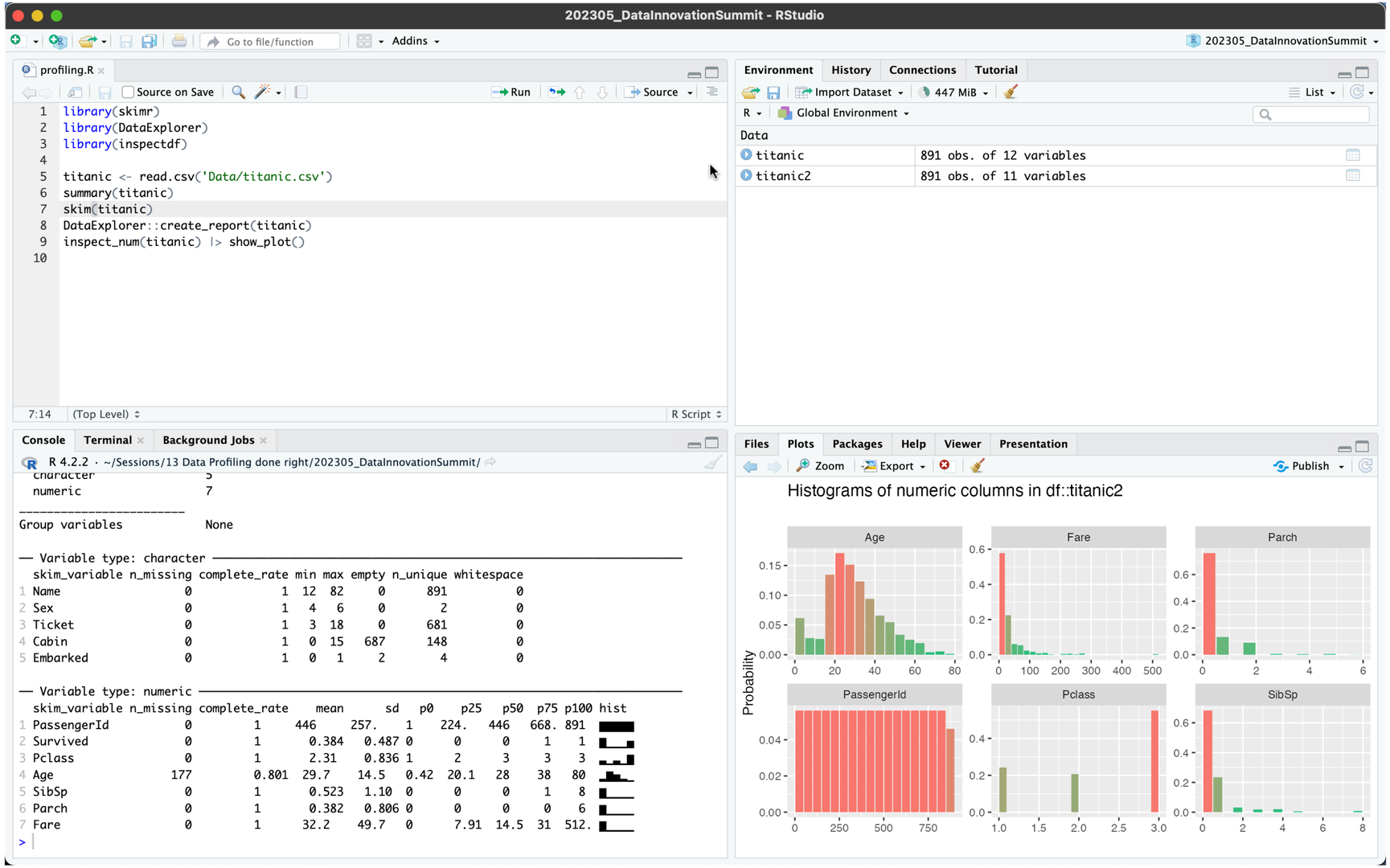




Techniques and tools: Excel / Power Query



Techniques and tools: R / RStudio



Techniques and tools: R / RStudio

```
> summary(titanic)
```

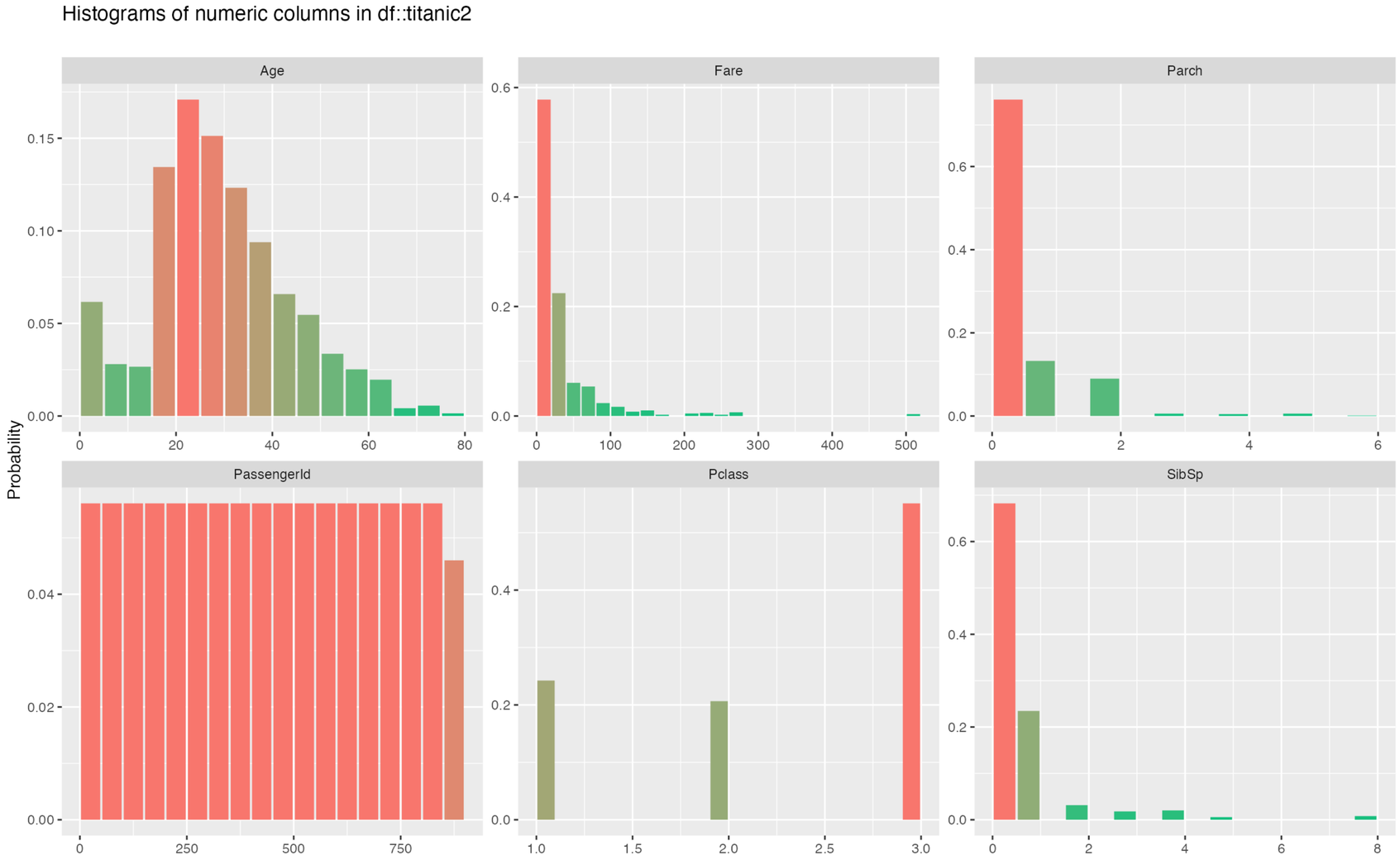
```
  PassengerId       Survived          Pclass          Name               Sex                 Age              SibSp       
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891         Length:891         Min.   : 0.42   Min.   :0.000  
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character   1st Qu.:20.12   1st Qu.:0.000  
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character   Mode  :character   Median :28.00   Median :0.000  
 Mean   :446.0   Mean   :0.3838   Mean   :2.309                                         Mean   :29.70   Mean   :0.523  
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                                         3rd Qu.:38.00   3rd Qu.:1.000  
 Max.   :891.0   Max.   :1.0000   Max.   :3.000                                         Max.   :80.00   Max.   :8.000  
                                                                                        NA's   :177                    
     Parch           Ticket               Fare            Cabin              Embarked        
 Min.   :0.0000   Length:891         Min.   :  0.00   Length:891         Length:891        
 1st Qu.:0.0000   Class :character   1st Qu.:  7.91   Class :character   Class :character  
 Median :0.0000   Mode  :character   Median : 14.45   Mode  :character   Mode  :character  
 Mean   :0.3816                      Mean   : 32.20                                        
 3rd Qu.:0.0000                      3rd Qu.: 31.00                                        
 Max.   :6.0000                      Max.   :512.33                                        
```

Techniques and tools: R / RStudio

```
> skim(titanic)
── Data Summary ────────────────────────
                           Values
Name                       titanic
Number of rows             891
Number of columns          12
                           ________________________
                           Column type frequency:
                             character               5
                             numeric                 7
── Variable type: character ────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 Name                  0             1  12  82     0      891          0
2 Sex                   0             1   4   6     0        2          0
3 Ticket                0             1   3  18     0      681          0
4 Cabin                 0             1   0  15   687      148          0
5 Embarked              0             1   0   1     2        4          0
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate    mean      sd   p0    p25   p50  p75 p100 hist
1 PassengerId           0             1     446     257.    1    224.   446   668. 891 ▇▇▇▇▇
2 Survived              0             1       0.384   0.487 0      0      0     1    1 ▇▁▁▁▅
3 Pclass                0             1       2.31    0.836 1      2      3     3    3 ▃▁▂▁▇
4 Age                 177             0.801  29.7    14.5   0.42  20.1   28    38   80 ▂▇▅▂▁
5 SibSp                 0             1       0.523   1.10  0      0      0     1    8 ▇▁▁▁▁
6 Parch                 0             1       0.382   0.806 0      0      0     0    6 ▇▁▁▁▁
7 Fare                  0             1      32.2    49.7   0      7.91  14.5  31   512. ▇▁▁▁▁
```

Techniques and tools: R / RStudio

```
> inspect_num(titanic2) |> show_plot()
```

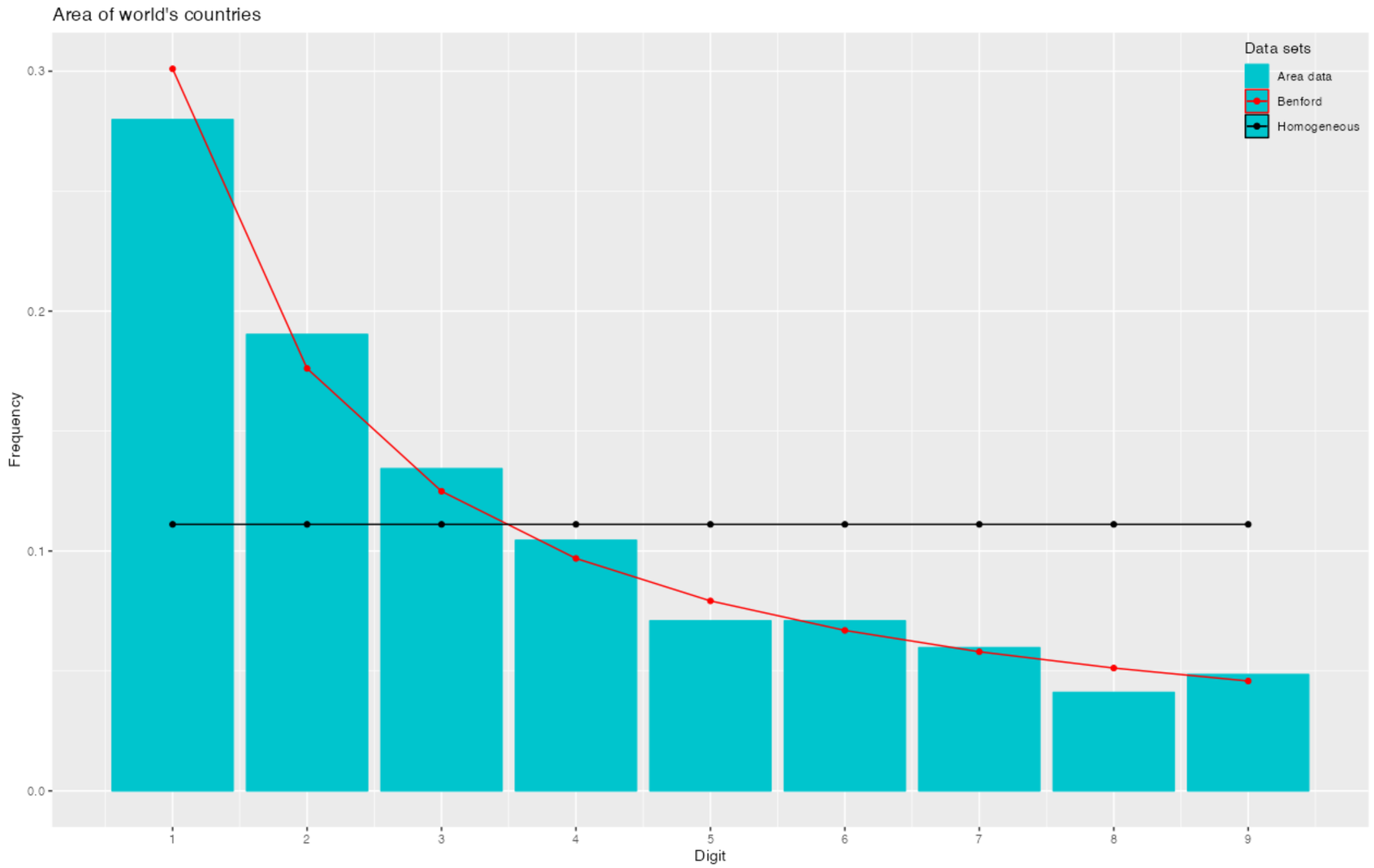


One more visualization example: Benford’s law

- is also known as „Newcomb-Benford-law“ or „Law of the first digit“
- is an observation about the distribution of leading digits in numerical data, first intuition (= equal distribution?) doesn’t always hold true
- applies to natural physical / mathematical or transactional data, the bigger the sample the better, spanning several orders of magnitude
- does not work for numbers influenced by human rules or actions
- has applications in accounting/economics (fraud detection), engineering, environment sciences, medicine, social sciences, election forensics, statistics…

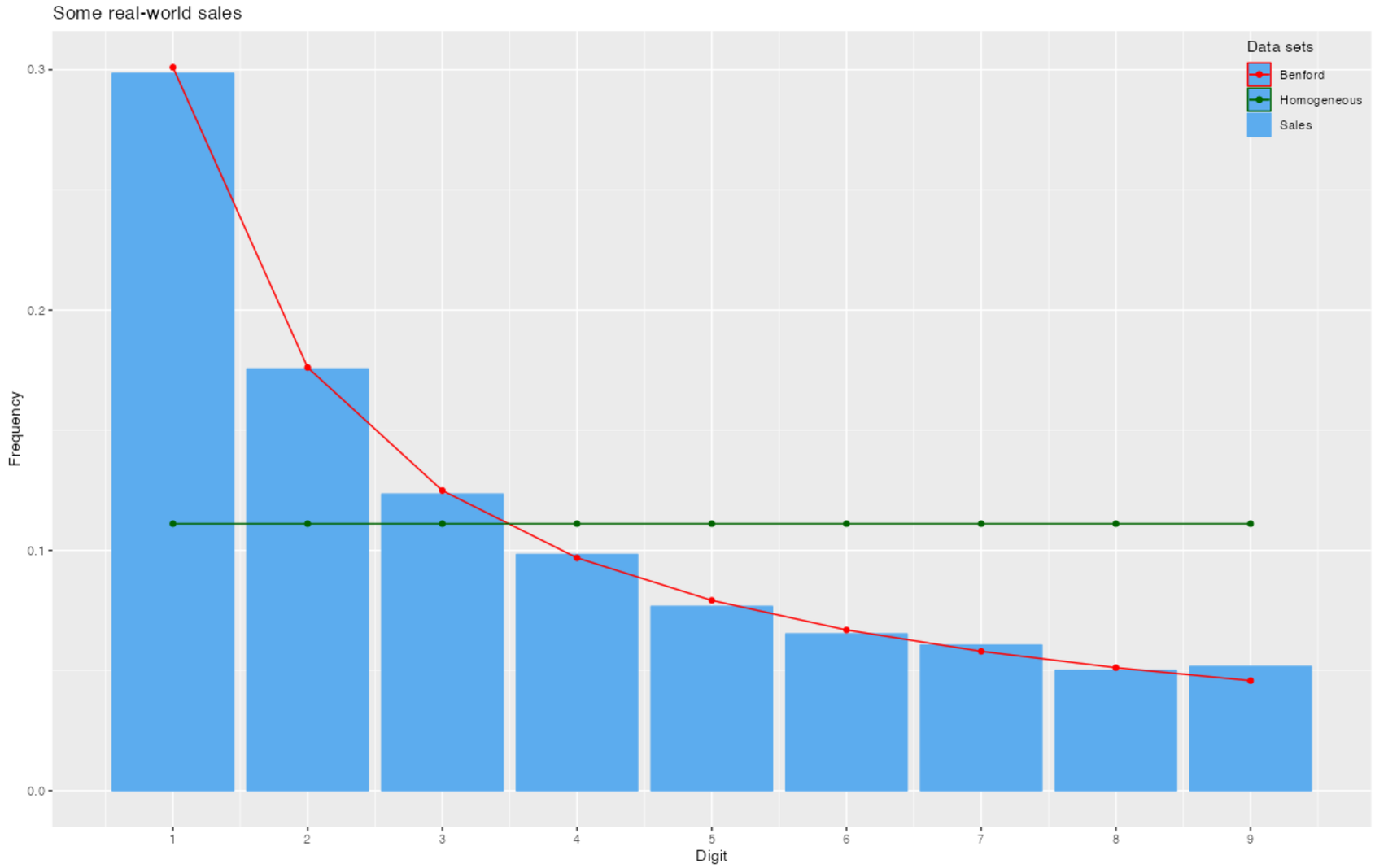
One more visualization example: Benford’s law

- Area of the world’s Countries



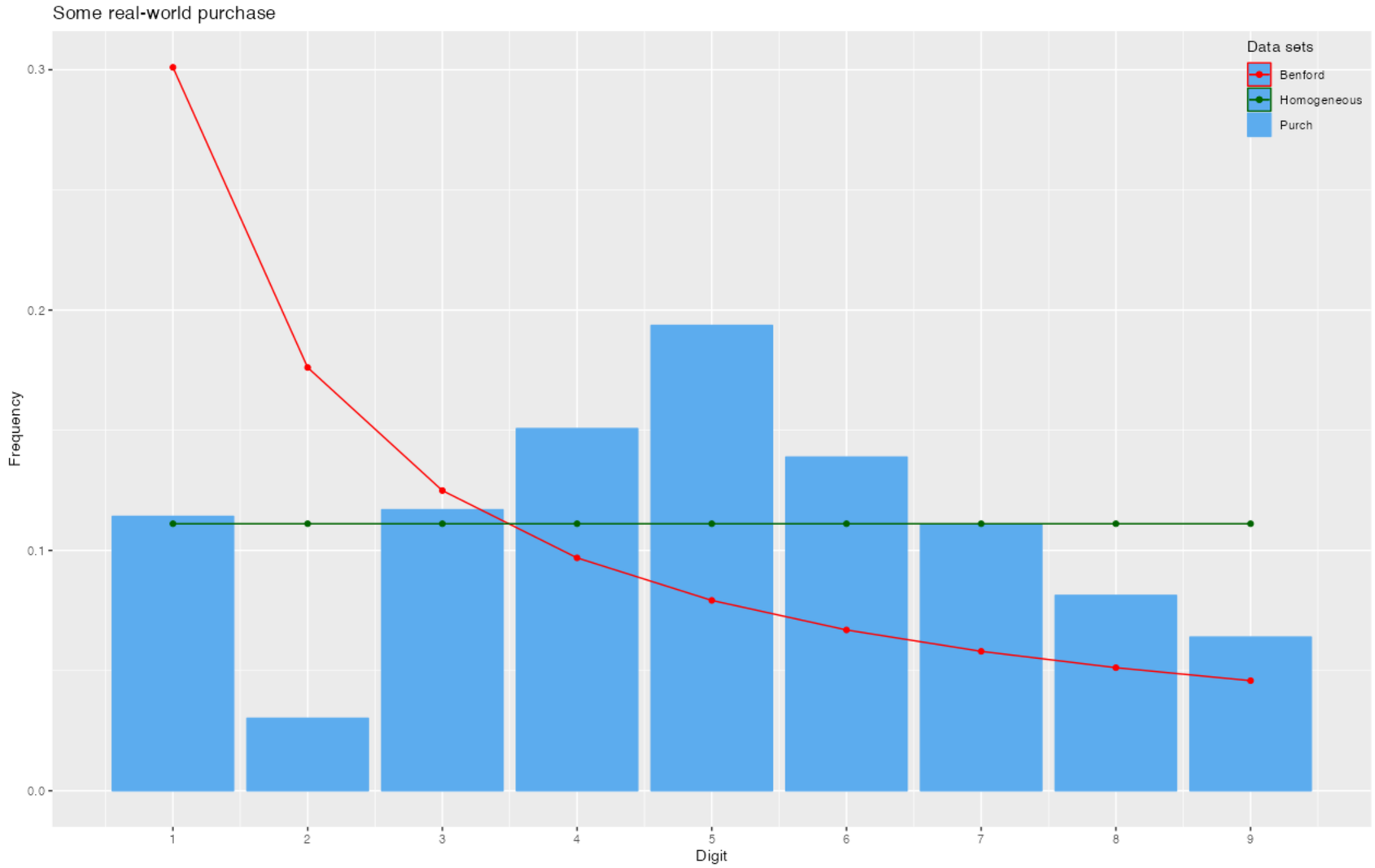
One more visualization example: Benford's law

- Sales invoices



One more visualization example: Benford's law

- Purchase invoices



Round-up / conclusions:

- For optimum performance of your business, keep your data quality high by implementing data profiling and cleansing processes and techniques.

- Keep your costs low by applying those measures as early as possible. Think of the 1-10-100 rule, or even worse, loss of reputation.

- Also true in data profiling: a picture says more than a thousand words

➡ Visualize your data!

Resources:

- Wikipedia: https://en.wikipedia.org/wiki/Data_profiling
- Gartner: https://www.gartner.com/en/information-technology/glossary/data-profiling
- Dr. Kirk Borne: https://medium.com/codex/data-profiling-having-that-first-date-with-your-data-2e05de50fca7
- Database normalization: https://en.wikipedia.org/wiki/Database_normalization, https://en.wikipedia.org/wiki/Codd%27s_12_rules
- HBR: https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year
- Data profiling tools: https://dbmstools.com/categories/data-profiling-tools
- Benford’s law: en.wikipedia.org/wiki/Benford's_law

Data profiling done right from the start

Thank you for your time and interest & keep in touch:

- @DerFredo https://twitter.com/DerFredo
- de.linkedin.com/in/derfredo
- https://techhub.social/@DerFredo

This file can be downloaded from:
https://bit.ly/DerFredoDIS23