Data Collection

TYPES OF DATA SOURCES

- More data is better to start off the analysis.
- Data can originate from a variety of different sources.
- **Transactions** are the first important source of data.
- Transactional data consist of structured, low-level, detailed information capturing the key characteristics of a customer transaction (e.g., purchase, claim, cash transfer, credit card payment).
- This type of data is usually stored in massive online transaction processing (OLTP) relational databases.
- It can also be summarized over longer time horizons by aggregating it into averages, absolute/relative trends, maximum/minimum values, and so on.

- Unstructured data embedded in text documents (e.g., emails, web pages, claim forms) or multimedia content can also be interesting to analyze.
- However, these sources typically require extensive preprocessing before they can be successfully included in an analytical exercise.

- Another important source of data is qualitative, **expert-based data**. An expert is a person with a substantial amount of subject matter expertise within a particular setting (e.g., credit portfolio manager, brand manager).
- The expertise stems from both common sense and business experience, and it is important to elicit expertise as much as possible before the analytics is run.
- This will steer the modeling in the right direction and allow you to interpret the analytical results from the right perspective.

- A popular example of applying expert-based validation is checking the univariate signs of a regression model.
- For example, one would expect *a priori* that higher debt has an adverse impact on credit risk, such that it should have a negative sign in the final scorecard.
- If this turns out not to be the case (e.g., due to bad data quality, multicollinearity), the expert/business user will not be tempted to use the analytical model at all, since it contradicts prior expectations.

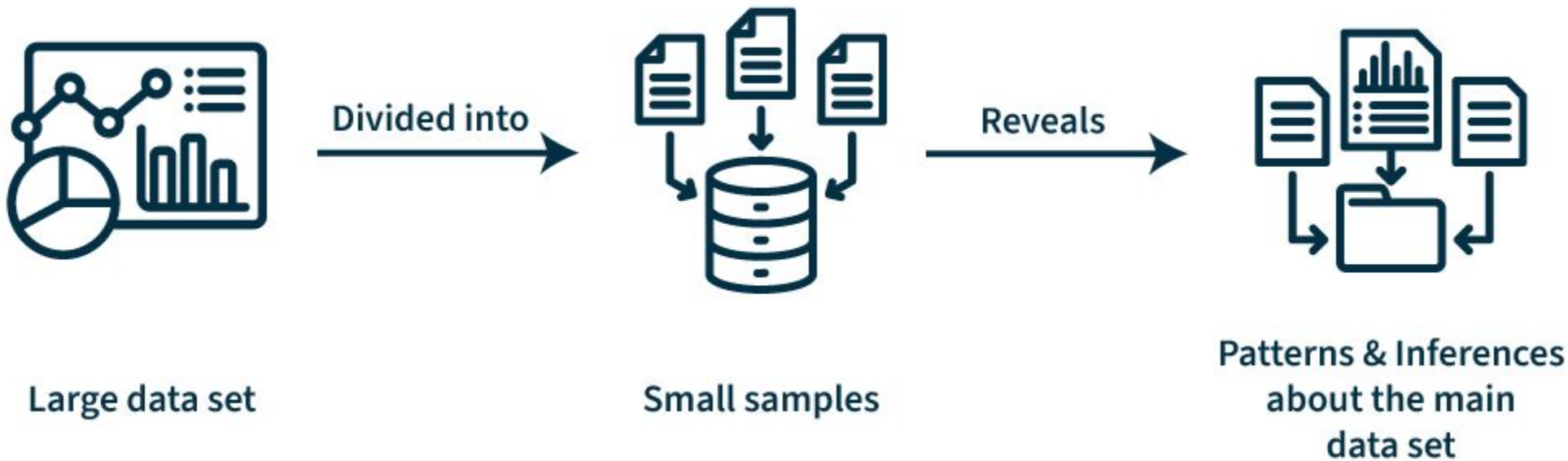
- Nowadays, **data poolers** play an important role in industry analytics.
- Data poolers are companies that collect large amounts of data in a specific domain, build analytical models on it, and sell the results, such as scores or reports, to other organizations.
- Popular examples include Dun & Bradstreet, Bureau van Dijk, and Thomson Reuters.
  They mainly operate in areas like credit risk, finance, and marketing.
- A well-known example is the FICO Score in the United States.
  It is a credit score between 300 and 850, provided by major credit bureaus such as Experian, Equifax, and TransUnion.

- Banks and financial institutions use these scores either as:
- Their main credit decision model, or A benchmark to compare with their own internal credit scorecards
- This helps them evaluate risk and identify weaknesses in their internal models.

- Finally, plenty of **publicly available data** can be included in the analytical exercise.
- A first important example is macroeconomic data about gross domestic product (GDP), inflation, unemployment, and so on.
- By including this type of data in an analytical model, it will become possible to see how the model varies with the state of the economy.
- This is especially relevant in a credit risk setting, where typically all models need to be thoroughly stress tested.
- In addition, social media data from Facebook, Twitter, and others can be an important source of information.
- However, one needs to be careful here and make sure that all data gathering respects both local and international privacy regulations.

SAMPLING

Sampling


Divided into
Reveals
Large data set
Small samples
Patterns & Inferences about the main data set

What is sampling?

- Sampling is the process of selecting a subset (sample) from a larger dataset (population) to analyze and draw conclusions about the entire population.

Why Sampling is Important (Relevance)

- Cost and Time Efficiency
  - Analyzing the entire population can be expensive and time-consuming.
  - Sampling reduces data collection, storage, and processing costs.
- Feasibility with Large Datasets
  - In big data scenarios (millions of records), full analysis may be impractical.
  - Sampling enables quick insights without handling the entire dataset.
- Faster Decision Making
  - Businesses and researchers can make timely decisions using sample-based analysis.

- Resource Optimization
  - Reduces computational load on systems (CPU, memory).
  - Enables analytics even with limited infrastructure.
- Accuracy with Proper Design
  - A well-chosen representative sample can provide results nearly as accurate as population analysis.
- Data Quality Improvement
  - Smaller datasets are easier to clean, validate, and explore.
  - Helps in early detection of trends, anomalies, or errors.
- Foundation for Statistical Inference
  - Sampling allows estimation of population parameters (mean, variance, proportion).
  - Forms the basis of hypothesis testing, confidence intervals, and predictive modeling.

Sampling Techniques in Data Analytics

- Sampling techniques are broadly classified into: probability and non-probability sampling.

Probability Sampling Techniques

Each element has a known, non-zero chance of selection

a) Simple Random Sampling

- Every data point has an equal chance of being selected.
- Methods: random number tables, random generators.
- Advantage: Unbiased, easy to understand.
- Limitation: Not ideal for very large or heterogeneous datasets.
- Example: Randomly selecting 500 patient records from a hospital database

b) Systematic Sampling

- Select every *k*th element after a random start.
- *k = population size / sample size*
- Advantage: Simple and fast.
- Limitation: Can be biased if data has periodic patterns.
- Example: Selecting every 10th transaction record.

c) Stratified Sampling

- Population divided into homogeneous groups (strata).
- Samples drawn from each stratum.
- Advantage: Ensures representation of all subgroups.
- Limitation: Requires prior knowledge of strata.
- Example: Sampling students separately from UG, PG, and PhD groups.

d) Cluster Sampling

- Population divided into clusters (heterogeneous groups).
- Entire clusters are randomly selected.
- Advantage: Cost-effective for geographically spread data.
- Limitation: Less accurate than stratified sampling.
- Example: Selecting a few colleges and surveying all students in them.

Non-Probability Sampling Techniques

a) Convenience Sampling

- Samples chosen based on ease of access.
- Advantage: Quick and inexpensive.
- Limitation: High bias, poor generalization.
- Example: Surveying people available in a classroom.

b) Judgment (Purposive) Sampling

- Researcher selects samples based on expertise.
- Advantage: Useful for expert-driven studies.
- Limitation: Subjective and biased.
- Example: Selecting experienced radiologists for medical image analysis validation.

c) Quota Sampling

- Population divided into groups; fixed quota from each group.
- Advantage: Ensures subgroup representation.
- Limitation: Non-random selection within groups.

d) Snowball Sampling

- Existing samples recruit future samples.
- Advantage: Useful for hidden or rare populations.
- Limitation: Dependency bias.
- Example: Identifying patients with rare genetic disorders.

TYPES OF DATA ELEMENTS

- It is important to appropriately consider the different types of data elements at the start of the analysis.
- The following types of data elements can be considered:
  - **Continuous**: These are data elements that are defined on an interval that can be limited or unlimited.
  - Examples : income, sales, RFM (recency, frequency, monetary).
  - **Categorical**
    - **Nominal:** These are data elements that can only take on a limited set of values with no meaningful ordering in between. Examples : marital status, profession, purpose of loan.
    - **Ordinal**: These are data elements that can only take on a limited set of values with a meaningful ordering in between. Examples : credit rating; age coded as young, middle aged, and old.
    - **Binary:** These are data elements that can only take on two values. Examples include gender, employment status.

Visual data exploration and exploratory statistical analysis

- Visual data exploration is a very important part of getting to know your data in an “informal” way.
- It allows you to get some initial insights into the data, which can then be usefully adopted throughout the modeling.
- Different plots/graphs can be useful here.
- example is pie charts, Bar charts, histograms and scatter plots
- A next step after visual analysis could be inspecting some basic statistical measurements, such as averages, standard deviations, minimum, maximum, percentiles, and confidence intervals.
- One could calculate these measures separately for each of the target classes (e.g., good versus bad customer) to see whether there are any interesting patterns present (e.g., whether bad payers usually have a lower average age than good payers).

Missing values

- Reasons:
  - The information can be nonapplicable.
  - The information can also be undisclosed
  - error during merging
- Some analytical techniques (e.g., decision trees) can directly deal with missing values.
- Other techniques need some additional preprocessing.
  - Various Schemes:
    - Replace (impute)
    - Delete
    - Keep

Table 2.1 Dealing with Missing Values

ID	Age	Income	Marital Status	Credit Bureau Score	Class
1	34	1,800	?	620	Churner
2	28	1,200	Single	?	Nonchurner
3	22	1,000	Single	?	Nonchurner
4	60	2,200	Widowed	700	Churner
5	58	2,000	Married	?	Nonchurner
6	44	?	?	?	Nonchurner
7	22	1,200	Single	?	Nonchurner
8	26	1,500	Married	350	Nonchurner
9	34	?	Single	?	Churner
10	50	2,100	Divorced	?	Nonchurner

- Replace (impute)
  - replacing the missing value with a known value
- Delete
  - most straightforward option and consists of deleting observations or variables with lots of missing values.
  - assumes that information is missing at random and has no meaningful interpretation and/or relationship to the target
- Keep
  - Missing values can be meaningful (e.g., a customer did not disclose his or her income because he or she is currently unemployed).

Outlier detection and treatment

- Outliers are extreme observations that are very dissimilar to the rest of the population.
  - Valid observations (e.g., salary of boss is $1 million)
  - Invalid observations (e.g., age is 300 years)
- Both are univariate outliers in the sense that they are outlying on one dimension.
- Multivariate outliers are observations that are outlying in multiple dimensions.
- Two important steps in dealing with outliers are detection and treatment.
- A first obvious check for outliers is to calculate the minimum and maximum values for each of the data elements.

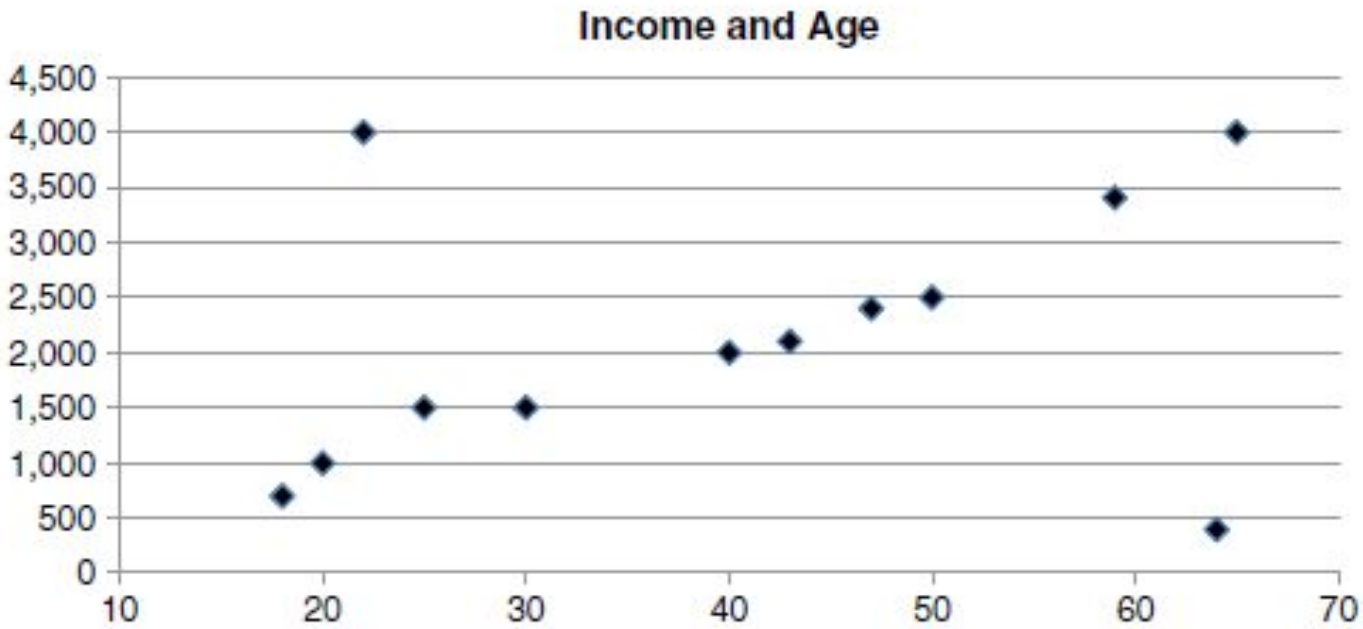
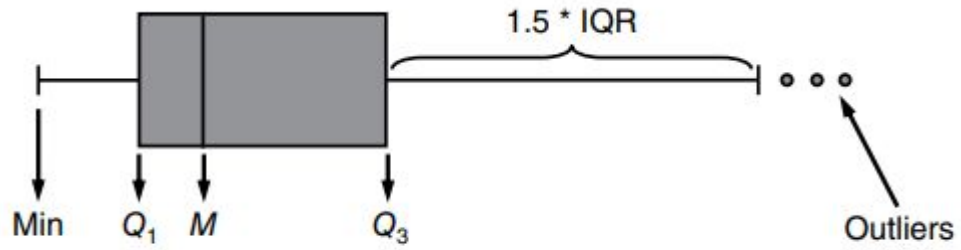


Figure 2.3 Multivariate Outliers

- Visual Mechanisms:
  - Histograms
  - Box plots
- Box Plots:
- A box plot represents three key quartiles of the data: the first quartile (25 percent of the observations have a lower value), the median (50 percent of the observations have a lower value), and the third quartile(75 percent of the observations have a lower value).
- All three quartiles are represented as a box.
- The minimum and maximum values are then also added unless they are too far away from the edges of the box.

- Too far away is then quantifi ed as more than 1.5 * Interquartile Range (IQR = Q 3 − Q1).



; Box Plots for Outlier Detection

- Lower limit = $Q1-1.5\times IQR$, Upper limit = $Q3+1.5\times IQR$.
- Any value ,below the lower limit or above the upper limit is considered an outlier.

- Another way is to calculate z-scores, measuring how many standard deviations an observation lies away from the mean, as follows:

$$z_i = \frac{x_i - \mu}{\sigma}$$

- Rule of thumb
- $| z | > 3 \rightarrow$ potential outlier

STANDARDIZING DATA

- Standardizing data is a data preprocessing activity targeted at scaling variables to a similar range.
- Min/max standardization

$$X_{new} = \frac{X_{old} - \min(X_{old})}{\max(X_{old}) - \min(X_{old})}(newmax - newmin) + newmin,$$

  - whereby newmax and newmin are the newly imposed maximum and minimum (e.g., 1 and 0)
- Z-score standardization
- Decimal scaling
  - Dividing by a power of 10 as follows:

$$X_{new} = \frac{X_{old}}{10^n},$$

  - with n the number of digits of the maximum absolute value.(largest value in the dataset , ignoring the sign).

Dimensionality Reduction

- Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results.

- Dimensionality reduction is a data reduction technique that reduces the number of input variables (attributes or features) in a dataset while preserving the essential information and structure of the data.

- It transforms high-dimensional data into a lower-dimensional representation that is easier to store, process, and analyze.

Why Dimensionality Reduction is Needed

- **Curse of Dimensionality**
- As the number of dimensions increases:
  - Data becomes sparse
  - Distance measures become less meaningful
  - Model performance degrades
- **Improved Computational Efficiency**
- High-dimensional data requires more memory and computation
- Training ML models becomes slow and expensive
- Reducing dimensions **speeds up algorithms** and reduces storage cost.

- **Noise and Redundancy Removal**
  - Real-world datasets contain irrelevant, noisy, and correlated features
  - These features reduce model accuracy
  - Dimensionality reduction removes redundant attributes and improves signal-to-noise ratio.
- **Visualization and Interpretability**
  - Humans cannot visualize data beyond 3 dimensions
  - Dimensionality reduction enables 2D or 3D visualization
  - Useful for data exploration and pattern discovery.

Types of Dimensionality Reduction

- **Feature Selection**
  - Selects a subset of original attributes
  - Example: Removing irrelevant features
- **Feature Extraction**
  - Creates new features by transforming original ones
  - Example: PCA, Wavelet Transform

PCA

- PCA transforms the original data into a new coordinate system such that:
- The first principal component has maximum variance
- Each subsequent component has the next highest variance
- Components are orthogonal (uncorrelated)
- Assumptions and Data Representation
  - The dataset consists of tuples (data objects)
  - Each tuple is represented as an n-dimensional vector
  - $X = (x_1, x_2, \ldots, x_n)$
  - Attributes are numeric and continuous
  - Attributes may be correlated

Working steps

1. Standardize or normalize the data , by calculating mean and standard deviation.

2. Next PCA calculates the covariance matrix to see how features relate to each other whether they increase or decrease together.

3. PCA identifies new axes where the data spreads out the most:

- 1st Principal Component (PC1): The direction of maximum variance (most spread).
- 2nd Principal Component (PC2): The next best direction, perpendicular to PC1 and so on.
- These directions come from the eigenvectors of the covariance matrix and their importance is measured by eigenvalues.

4.Pick the Top Directions (Select top k eigenvectors)

- After calculating the eigenvalues and eigenvectors PCA ranks them by the amount of information they capture. We then:
- Select the top k components that capture most of the variance like 95%.
- Transform the original dataset by projecting it onto these top components.

5. Transform data

- Project data onto the new k-dimensional space.
- This means we reduce the number of features (dimensions) while keeping the important patterns in the data.

Wavelet Transforms

- **Principle**
- Wavelet Transform represents data using **wavelet coefficients** at different scales and resolutions.
- The Wavelet Transform (WT) reduces dimensionality by converting data into wavelet coefficients, then thresholding (setting small coefficients to zero), creating a sparse representation that retains key features while drastically cutting data size, making it ideal for signal processing, image compression, and large datasets by concentrating energy/information into fewer significant values

Working

- **Transformation**: The Discrete Wavelet Transform (DWT) converts an input data vector (e.g., signal, image row) into a set of wavelet coefficients representing different scales (frequencies) and locations.
- **Sparsification**: A threshold is applied; coefficients below this threshold are set to zero. This highlights significant data features (e.g., edges in images, sudden changes in signals) while discarding less important "noise".
- **Representation**: The result is a sparse dataset (many zeros) with fewer non-zero coefficients, effectively reducing dimensions.
- **Reconstruction (Optional)**: An inverse DWT can reconstruct an approximation of the original data from the reduced set of coefficients, which is useful for compression.