3.1 Data Collection, Sampling, and Preprocessing

3.1.1 Real Data #RealData

In theory: *"The bigger the better"* but **real data is (typically) "dirty"**:

- Inconsistencies.
- Incompleteness.
- Duplication.
- ...

> *"Messy data will yield messy analytical models"*

Data-filtering mechanisms applied to *clean up and reduce the data*.

Even **the slightest mistake** can make the *data totally unusable and the results invalid*.

3.1.2 Types of Data Sources #DataSources

Variety of **different sources** that *provide* **different types of information**:

- Transactional data.
- Contractual, subscription, or account data.
- Sociodemographic information.
- Surveys.
- Behavioral information.
- Unstructured data.
- Contextual or network information.
- Qualitative, expert-based data.
- Publicly available data.
- ...

3.1.2.1 Transactional data #TransactionalData

#DEF **Structured and detailed information capturing the key characteristics of a customer transaction.**

Summarized over longer time horizons *by aggregating it*:

- Averages.
- (Absolute or relative) trends.
- Maximum or minimum values.
- Recency (R), Frequency (F), and Monetary (M).

> *Meaningful when interpreted individually*
> *Their interaction is very useful for fraud detection, anti-money laundering.*

3.1.3 Types of Data Elements #DataElements

- **Continuous data**: *Data elements defined on an interval*, which can be both *limited and unlimited*.
- **Categorical data**:
  - *Nominal*: data elements that can only take on a limited set of values with **no** meaningful **ordering** in between.
  - *Ordinal*: data elements that can only take on a limited set of values **with** a **meaningful ordering** in between.
  - *Binary*: data elements that can only take **two values** (yes/no).

3.1.4 Sampling #Sampling

Take a subset of historical data to build an analytical model.

❓ Why not analyze directly the full data set?
🔑 **Key requirement for a good sample** = *representative for the future entities*.

> *Timing and representativeness are crucial!*

3.1.4.1 Sampling Timing and Bias

Choosing the **optimal time window** is a **trade-off** between:

- *Lots of data* (a more robust analytical model).
- *Recent data* (more representative).

An **"average" period** to get *as accurate as possible* a picture of the *target population*.

> *Sampling bias should be avoided even if not straightforward.*

Bias: Credit card context Example #Bias

Scenario: Customers may use their credit card differently during the *month of December* when *buying gifts for the holiday period*.

Two sources of bias from normal business periods:

1. Credit card customers may *spend more during this period*, both in total as well as on individual products.
2. *Different types of products* may be bought in different stores usually frequented by the customer.

3.1.4.2 Mitigations to address seasonality effect or bias #Bias

Every month may deviate from normal (i.e., average):

1. **Build separate models for different months, or for homogeneous time frames**:
   This is a *complex and demanding* solution: multiple models have to be developed, run, maintained, and monitored.
2. **Sampling observations over a period covering a full business cycle and build a single model**:
   - Cost of *reduced fraud detection power* since *less tailored to a particular time frame*.
   - *Less complex and costly* to operate.

> ***Sampling** has a direct impact on the fraud detection power.*

3.1.4.3 Stratified Sampling

#DEF **A sample is taken according to predefined strata.**

In a fraud detection context *data sets are very skew*.
Stratifying according to the target fraud indicator:

- *Sample will contain exactly the same percentages* of (non) fraudulent transactions *as in the original data*.
  Stratification applied on predictor variables:
- *Resemble* the real product *transaction distribution*.

3.1.5 Exploratory Statistical Analysis #StatisticalAnalysis

Inspect some basic statistical measurements:

- Averages.
- Standard deviations.
- Minimum, maximum.
- Percentiles.
- Confidence intervals.
- ...

> *Calculate these measures separately for each of the target classes (e.g., fraudsters versus non fraudsters) to see whether there are any interesting patterns present.*

3.1.5.1 Basic Descriptive Statistics

Descriptive statistics provide basic insight for the data.
They should be *assessed together* (in support and completion of each other):

- The *mean and median* value of *continuous variables*:

> - The *median value* less sensitive to extreme values but not provide as much information with respect to the full distribution.

- The *variation or the standard deviation* provide insight with respect to how much the data is spread around the mean value.
- *Percentile values*, provide complementary information w.r.t. the distribution and the median value.
- With categorical variables, one may calculate the *mode*, which is the most frequently occurring value.

3.1.5.2 Specific Descriptive Statistics

Express the symmetry or asymmetry of a distribution (e.g., skewness, peakedness or flatness of a distribution).

The values of these measures are harder to interpret:

- *Limits* their practical use.
- Sometime it is *easier* to assess these aspects *by inspecting visual plots of the distributions of the involved variables*.

3.1.6 Missing Values #MissingValues

Missing values can occur because of various reasons:

- The *information* can be *non applicable*.
- The *information* can also be *undisclosed*.
- *Error* during merging.

> *Some **analytical techniques** (e.g., decision trees) can deal directly with missing values. Other techniques need some additional preprocessing.*

3.1.6.1 Dealing with Missing Values

- **Replace**: replacing the *missing value with a known value*.
- **Delete**: *deleting observations or variables with lots of missing values*. This assumes that information is missing at random and has no meaningful interpretation and/or relationship to the target.
- **Keep** *Missing values can be meaningful* and may have a *relation with fraud* and needs to be considered as a separate category.

Statistically test whether *missing information is related to the target variable or not.*

- **If yes**, then we can *adopt the keep strategy and make a special category for it*.
- **If not**, one can depending on the number of observations available, decide to *either delete or replace*.

3.1.7 Outliers #Outliers

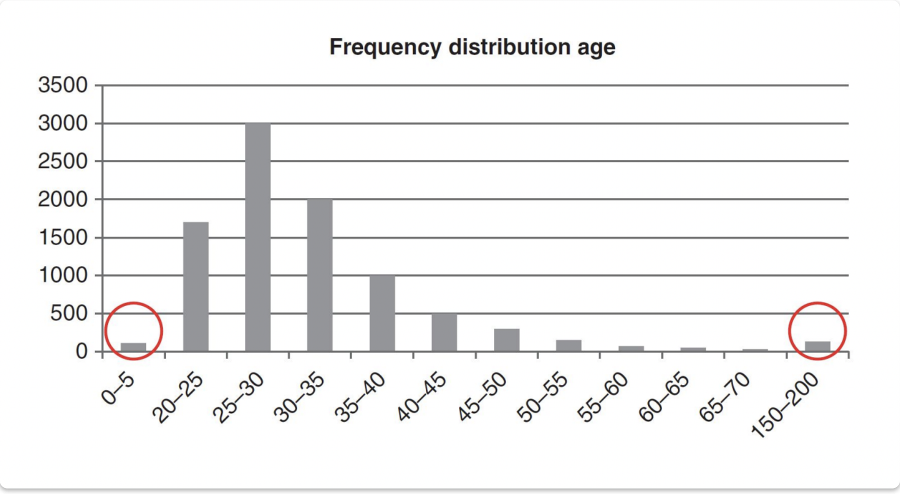
#DEF **Extreme observations that are very dissimilar to the rest of the population.**

- *Valid observations*: e.g., salary of boss is US$1,000,000.
- *Invalid observations*: e.g., age is 300 years.
- **Univariate outliers**: outlying on *one dimension*.
  Multivariate outliers: outlying in *multiple dimensions*.

3.1.7.1 Univariate Outlier Detection and Treatment

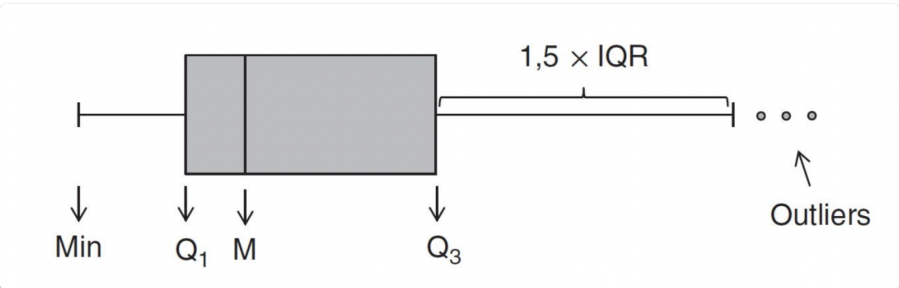
Minimum and maximum values for each of the data elements.
Graphical tools:

- *Histograms*:



- *Box Plot* [Univariate Variable]:
  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).



Z-scores [Univariate Variable]:
Measures how many standard deviations an observation lies away from the mean.

Z-score relies on the normal distribution.

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

ID	Age	z-Score
1	30	$(30 - 40)/10 = -1$
2	50	$(50 - 40)/10 = +1$
3	10	$\boxed{(10 - 40)/10 = -3}$
4	40	$(40 - 40)/10 = 0$
5	60	$(60 - 40)/10 = +2$
6	80	$\boxed{(80 - 40)/10 = +4}$
…		
	$\mu = 40$ $\sigma = 10$	$\mu = 0$ $\sigma = 1$

Fitting regression lines and inspecting the observations with large errors (using, e.g., a residual plot) [Multivariate Variable].

Clustering or calculating the Mahalanobis distance [Multivariate Variable].

3.1.7.2 Outlier Detection and Treatment

Various schemes exist to deal with outliers:

- For **invalid observations**, one could *treat the outlier as a missing value*.
- For **valid observations**: Impose both a *lower and upper limit* on a variable and *any values below/above are brought back to these limits*.

3.1.7.3 Expert-based limits based on business knowledge

#DEF **Not all invalid values are outlying and may go unnoticed if not explicitly looked into.**
Construct a *set of rules formulated based on expert knowledge*, which is applied to the data to check and alert for issues.

- Relations that exist between the different variables.
- Constraints that apply to the combination of variable values.

Example:
Customers:

- Birth date = "01/01/1980"

- Category = child
  Which value is *invalid* ? Cannot be determined...

  > *Both values are not outlying and therefore such a conflict will not be noted by the analyst unless some explicit precautions are taken.*

3.1.8 Discussion on preprocessing #Preprocessing

When handling valid outliers in the data set using the treatment techniques, *we may impair the ability of descriptive analytics in detecting frauds*:

- Be extremely careful in treating valid outliers *when applying unsupervised learning techniques to build a fraud detection model*.

When handling invalid outliers, on the contrary, *they can be treated as missing values* preferably by including an indicator that the value was missing *or even more precisely an invalid outlier*.

3.1.9 Standardizing Data #Standardize

#DEF **Scaling variables to a similar range.**

Example

- Gender (coded as 0/1).
- Income (ranging between 0 and US$1,000,000).
  Min/Max standardization: Whereby newmax and newmin are the newly imposed maximum and minimum (e.g., 1 and 0)
  Z-score standardization: Calculate the z-scores
  Decimal scaling

3.1.10 Categorization

#DEF **For categorical variables, it is needed to reduce the number of categories.** (E.g., IBAN, IP)

Basic methods:

- *Equal interval binning*: Bins with the same range.
- *Equal frequency binning*: Bins with the same number of observations.
- *Chi-squared analysis*
- *Pivot Table*

> *For **continuous variables**, by categorizing the variable into ranges, nonmonotonicity can be taken into account.*

3.1.11 Variable Selection #Variables

Many analytical modeling exercises **start with tons of variables**, of which *typically only a few actually contribute to the prediction of the target variable*.

> *The average model in fraud detection has between 10 and 15 variables.*

3.1.11.1 Filters #Filters

#DEF **Measure univariate correlations between each variable and the target.**

> *Are a very handy variable selection mechanism.*
>
> *Allow a quick screening of which variables should be retained for further analysis.*

	Continuous Target (e.g., CLV, LGD)	Categorical Target (e.g., churn, fraud, credit risk)
Continuous variable	Pearson correlation	Fisher score
Categorical variable	Fisher score/ANOVA	Information value Cramer's V Gain/entropy

3.1.11.2 Filtering discussion

👍 Advantages:
Filters allow *reduction in the number of dimensions* of the data set early in the analysis.

👎 Drawbacks:
Work univariately and **do not consider correlation between the dimensions individually**.

> *We need other criteria to further refine the characteristics.*
>
> - *Privacy issues and regulatory compliance.*
> - *Also operational issues could be considered.*

3.1.11.3 Principal Components Analysis #PCA

#DEF **Technique to reduce the dimensionality of data by forming new variables that are not correlated and linear composites of the original variables.**

These *new variables describe the main components or dimensions* that are *present in the original data set*.

> *Max number of new variables (i.e., the number of principal components) = number of original variables.*

The information (*variance*) contained in the set of original variables can be *summarized* by a *limited number of principal components*.
In theory, to explain all the variance in the original data set, the full set of principal components is needed.

Some of these only account for a very small fraction of variance of the original variables. Therefore, they can be left out.

PCA gives us: **reduced dimensionality in the data set.**

3.1.11.3.1 PCA LIMITATIONS: #PCALIMITATIONS

Replacing the original variables with a (reduced) set of uncorrelated principal components *comes at a price*:

- **Reduced interpretability.**

The principal component variables derived from the original set of variables *cannot easily be interpreted* = they are calculated as a weighted linear combination of the original variables.

When interpretability is no concern, then *PCA is a powerful data reduction tool that will yield a better model in terms of stability as well as predictive performance*.

3.1.12 Correlation and stability

#DEF **Stability or robustness of a model** = *stability of model's parameters estimated based on the observations*.

#DEF **Values of the parameters** = *relation between the explanatory or predictor variables and the dependent or target variable*.

🚧 Unstable Model:

- If the values of these *parameters heavily depend on the exact sample of observations used to induce the model*.
- *Correlation* among the explanatory or predictor variables (multicollinearity).

🕵️ Input selection procedure is often performed:

- *Filter* approach.
- A new set of factors may be derived using *principal component analysis* since the resulting new variables (i.e., *the principal components, will be uncorrelated among themselves*).

> *This gives us a **stable model!***

Next chapter: Descriptive Analytics for Fraud Detection