Affaires mondiales Canada | Global Affairs Canada

Canada

CANADIAN FOREIGN SERVICE INSTITUTE | L'INSTITUT CANADIEN DU SERVICE EXTÉRIEUR

Introduction to Data Analysis

DATA PROCESSING

Patrick Boily
Data Action Lab | uOttawa | Idlewyld Analytics
pboily@uottawa.ca

FOUR VERY IMPORTANT REMARKS

NEVER work on the original dataset. Make copies along the way.

Document **ALL** your cleaning steps and procedures.

If you find yourself cleaning too much of your data, **STOP**. Something might be off with the data collection procedure.

Think **TWICE** before discarding an entire record.

APPROACHES TO DATA CLEANING

There are two **philosophical** approaches to data cleaning and validation:

- methodical
- narrative

The **methodical** approach consists of running through a **check list** of potential issues and flagging those that apply to the data.

The **narrative** approach consists of **exploring** the dataset and trying to spot unlikely and irregular patterns.

Data Cleaning Bingo

random missing values	outliers	values outside of expected range - numeric	factors incorrectly/iconsistentl y coded	date/time values in multiple formats
impossible numeric values	leading or trailing white space	badly formatted date/time values	non-random missing values	logical inconsistencies across fields
characters in numeric field	values outside of expected range - date/time	DCB!	inconsistent or no distinction between null, 0,not available, not applicable,missing	possible factors missing
multiple symbols used for missing values	???	fields incorrectly separated in row	blank fields	logical iconsistencies within field
entire blank rows	character encoding issues	duplicate value in unique field	non-factor values in factor	numeric values in character field

TYPES OF MISSING OBSERVATIONS

Blank fields come in 4 flavours:

- **Nonresponse**
  an observation was expected but none had been entered
- **Data Entry Issue**
  an observation was recorded but was not entered in the dataset
- **Invalid Entry**
  an observation was recorded but was considered invalid and has been removed
- **Expected Blank**
  a field has been left blank, but expectedly so

Too many missing values (of the first three type) can be indicative of **issues with the data collection process** (more on this later); too many missing values (of the fourth type) can be indicative of **poor questionnaire design**.

THE CASE FOR IMPUTATION

Not all analytical methods can easily accommodate missing observations – 2 options:

- **Discard** the missing observation
  - not recommended, unless the data is missing completely randomly in the dataset as a whole
  - acceptable in certain situations (such as a small number of missing values in a large dataset)
- Come up with a **replacement (imputation) value**
  - main drawback: we never know what the true value would have been
  - often the best available option

TAKE-AWAYS

Missing values cannot simply be ignored.

The missing mechanism cannot typically be determined with any certainty.

Imputation methods work best when values are missing completely at random or missing at random, but imputation methods tend to produce biased estimates.

In single imputation, imputed data is treated as the actual data; multiple imputation can help reduce the noise.

Is stochastic imputation best? In our example, yes – but beware the *No-Free Lunch* theorem!

DETECTING ANOMALIES

Outliers may be anomalous along any of the unit's variables, or in combination.

Anomalies are by definition **infrequent**, and typically shrouded in **uncertainty** due to small sample sizes.

Differentiating anomalies from noise or data entry errors is **hard**.

Boundaries between normal and deviating units may be **fuzzy**.

When anomalies are associated with malicious activities, they are typically **disguised**.

DETECTING ANOMALIES

Numerous methods exist to identify anomalous observations; **none of them are foolproof** and judgement must be used.

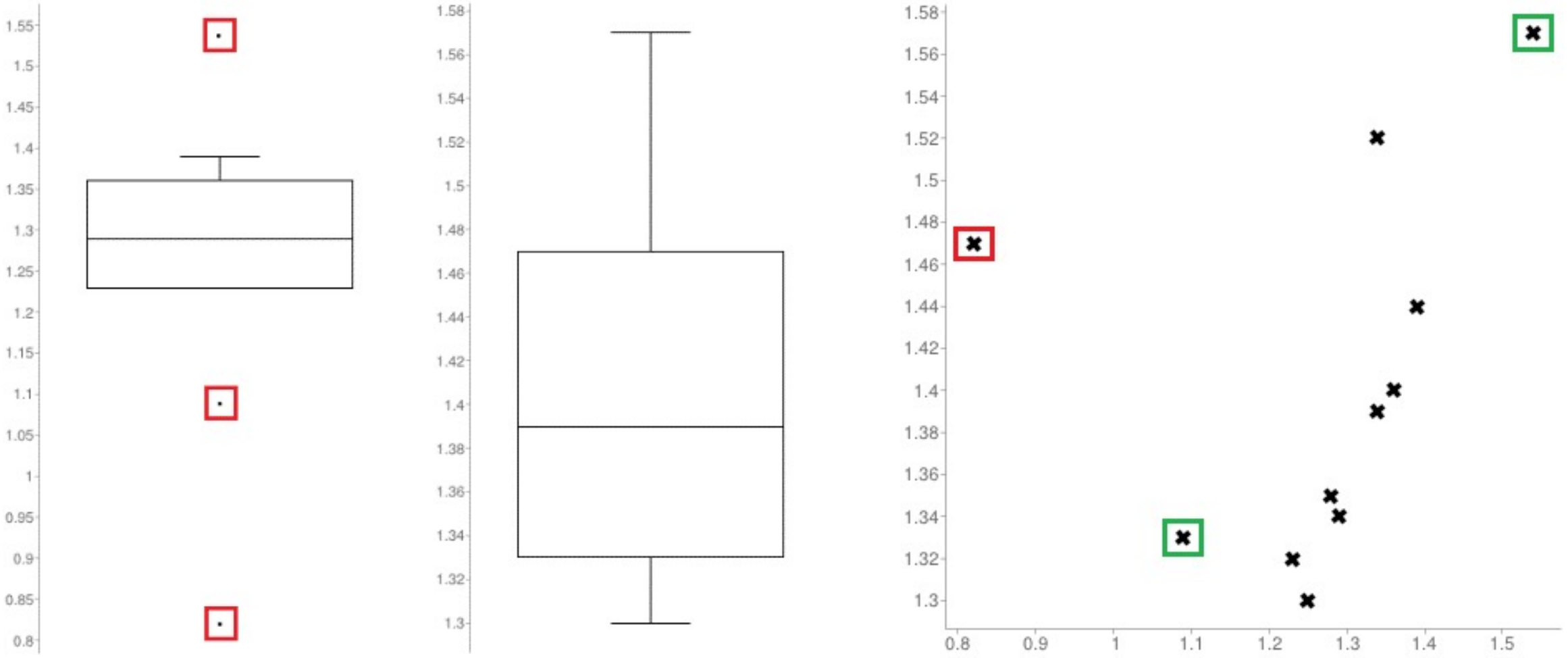
Graphical methods are easy to implement and interpret.

- **Outlying Observations**
  box-plots, scatterplots, scatterplot matrices, Cooke's distance, normal qq plots
- **Influential Data**
  some level of analysis must be performed (leverage)

Once anomalous observations have been removed from the dataset, previously “regular" units may become anomalous.

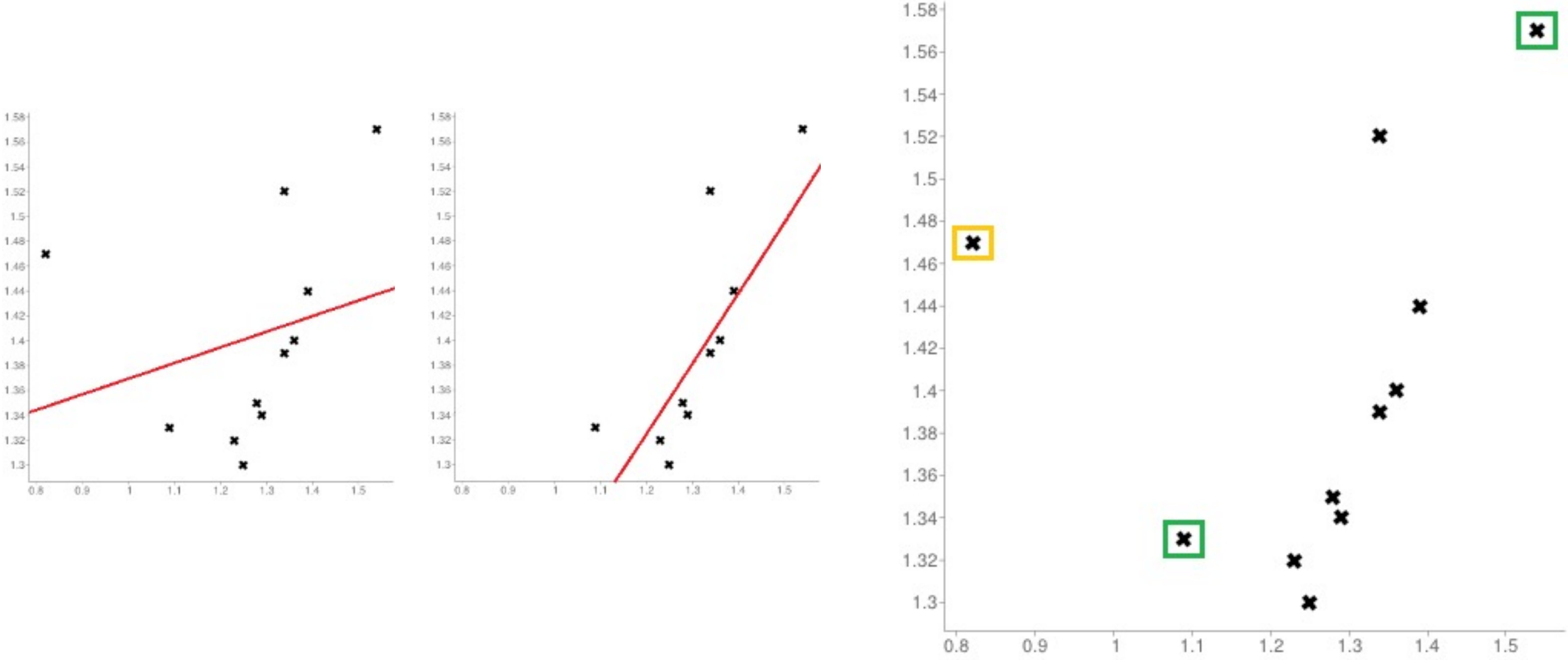
[Personal file]

OUTLIERS



Queuing dataset: processing rate vs. arrival rate

INFLUENTIAL OBSERVATIONS



Queuing dataset: processing rate vs. arrival rate

TAKE-AWAYS

Identifying influential points is an iterative process as the various analyses have to be run numerous times.

Fully automated identification and removal of anomalous observations is **NOT recommended**.

Use transformations if the data is **NOT** normally distributed.

Whether an observation is an outlier or not depends on various factors; what observations end up being influential data points depends on the specific analysis to be performed.

DIMENSIONALITY OF DATA

In data analysis, the **dimension** of the data is the number of variables (or attributes) that are collected in a dataset, represented by the number of columns.

The term dimension is an extension of the use of the term to refer to the size of a vector.

We can think of the variables used to describe each object (row) as a **vector** describing that object.

Note: the term dimension is used differently in business intelligence contexts.

CURSE OF DIMENSIONALITY

Unless the dataset size grows exponentially with its dimension, the performance of any model we build is likely to suffer due to the **Curse of Dimensionality**.

Possible solutions:

- **sampling observations**
- **feature selection** (easy-ish) and/or dimension reduction (hard).

We look for ways to preserve the signal while shrinking the dimension: it's easier to find needles in small haystacks!

(This is actually a thorny problem... but we'll avoid the technical details in this course).

FEATURE SELECTION

Removing **irrelevant** or **redundant** variables is a common data processing task.

Motivations:

- modeling tools do not handle these well (variance inflation due to multi-colinearity, etc.)
- dimension reduction (# variables $\gg$ # observations)

Approaches:

- filter vs. wrapper
- unsupervised vs. supervised

DISCRETIZING

To reduce computational complexity, a numeric variable may need to be replaced by an **ordinal** variable (from *height* value to "*short*", "*average*", "*tall*", for instance).

Domain expertise can be used to determine the bins' limits (although that could introduce unconscious bias to the analyses)

In the absence of such expertise, limits can be set so that either

- the bins each contain the same number of observations
- the bins each have the same width
- the performance of some modeling tool is maximized

SOUND DATA

The ideal dataset will have as few issues as possible with:

- **Validity:** data type, range, mandatory response, uniqueness, value, regular expressions
- **Completeness:** missing observations
- **Accuracy and Precision:** related to measurement and/or data entry errors; target diagrams (accuracy as bias, precision as standard error)
- **Consistency:** conflicting observations
- **Uniformity:** are units used uniformly throughout?

Checking for data quality issues at an early stage can save headaches later in the analysis.

SOUND DATA

accurate and precise

precise but not accurate

accurate but not precise

neither accurate nor very precise

TAKE-AWAYS

Don't wait until after the analysis to find out there was a problem with data quality.

Univariate tests don’t always tell the whole story.

Visualizations can help.

Context is crucial – you may need more context about the data in order to make sense of what you see… but whatever the situation, you need to understand the dataset quality.