COMP 333 — Week 7 Data Cleaning

Recap: Data Wrangling

Data wrangling is extremely important because your data is typically "messy"
and remember Garbage-In-Garbage-Out (GIGO) rule for computation
so you need to tidy-up your data before doing "serious" work.

Data wrangling is generally 60%+ of the time and effort for data analytics!

For data wrangling, you need to look closely at your data, so DDA is a basic tool.

Steps in data wrangling:
Step 1: Discover
Step 2: Structure
Step 3: Cleanse
Step 4: Enrich
Step 5: Validate
Step 6: Publish

Data wrangling is the traditional ETL (Extract-Load-Transform) process
from data warehouses and OLAP (online analytical processing).

In the data warehouse literature, the term *data cleansing*
is often used.
They take a more process-oriented approach there.
See the wikipedia entry
`https://en.wikipedia.org/wiki/Data_cleansing`.

Data Cleaning

Remember GIGO (Garbage-In, Garbage-Out)

Definition *Data cleaning* is the process of preparing data for analysis
by removing or modifying data that is
incorrect, incomplete, irrelevant, duplicated, or improperly formatted.
This data is usually not necessary or helpful when it comes to analyzing
data because it may hinder the process or provide inaccurate results.

A good article on data cleaning is
The Ultimate Guide to Data Cleaning: When the data is spewing garbage, by Omar Elgabry.
`https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4`

In this article, you can see a lot of influence from data warehousing
and a focus on process.
Data cleaning is often an *ad hoc* manual task.
However, when you deploy, as you must do for data warehouses,
the data cleaning steps need to be coded as a workflow.

Desirable Properties of Clean Data

The goal of data cleaning are to achieve *"good"* properties for the data.

- Validity
  The degree to which the data conform to defined business rules or constraints.
- Accuracy
  The degree to which the data is close to the true values.
- Completeness
  The degree to which all required data is known.
- Consistency
  The degree to which the data is consistent, within the same data set or across multiple data sets.
- Uniformity
  The degree to which the data is specified using the same unit of measure.

Workflow

The data cleaning workflow is an iteration of the following steps:

Inspection : Detect unexpected, incorrect, and inconsistent data.

Cleaning : Fix or remove the anomalies discovered.

Verifying : After cleaning, the results are inspected to verify correctness.

Reporting : Record a report about the changes made and the quality of the current data.

The inspection step is quantitative and visual Descriptive Data Analysis.

The cleaning step we will discuss at length.

For the verification step, when done with cleaning, you
should verify correctness by re-inspecting the data and
making sure the rules and constraints do hold.
That is, DDA again.

The reporting step should measure, plot, and report
how healthy the data is.
In addition to logging the violations,
the causes of these errors should be considered.
Why did they happen in the first place?

Track provenance, that is, the source of the data
and the reason for the errors

Issues with Data

Issues for data are:

- errors in data
- outliers and anomalies
- missing values and imputation of missing values
- duplicate data
- irrelevant data
- unification and normalization so data is comparable
- entity resolution

We discuss how to deal with each of these issues in the following lecture segments.

Recap: Advice from Data Wrangling Video

Advice from the video on Points-Of-Interest system (Week 6)

Data cleaning is an iterative process

Keep the raw data from data acquisition
You need a reference data set
you will need to re-process it many times

Script your cleaning steps
to refine and re-run your process

Recap: Advice from OpenRefine Video

Clustering of data values and clustering of observations
help you spot duplicates and "similar" values and observations
to make it easy to see differences/errors

Differences need to be investigated

In the case of large datasets,
be sure to limit your sample size
to see what data issues occur
before cleaning the whole dataset

Spot check throughout to prevent any errors from being replicated

Remember GIGO (Garbage-In, Garbage-Out)
There can be so much Garbage!

Skiena: Errors in Data

It is not so easy to know whether a value is an error, or not.
You want to distinguish between errors, artefacts (systematic errors)
and anomalies.

Error is information lost in data acquisition

Artifact systematic problem arising from data processing

Anomaly something that deviates from what is standard, normal, or expected
An anomaly may be a true value for your data,
or it may be an error.

And even *"expected"* values may be errors.

You will develop a *"nose"* for what is right or wrong with data
through inspection (DDA) and by returning to the data source itself
for answers to your doubts about the data.

This is called the **Sniff Test**:
look closely to see if something may be wrong

Outliers and Anomalies In everyday English,
there is no difference between outliers and anomalies.
The word *"anomaly"* is used also in the context of general anomalies
(not just in data or related to statistics).
The word *"outlier"* is used more exclusively to describe anomalies in data.

In data analytics, Tukey's definitions give us a computational way to decide on
outliers and extreme values.

Outliers

General advice is to regard *extreme values*
according to Tukey's definition, as true *outliers*

and to regard Tukey's *outlier values*
as potential outliers.

The ways to handle outliers are

- **keep** the observation with the outlier
- **drop** the observation with the outlier
- **flag** the outlier value

You should **flag** them

but by engineering a new feature/column *Outlier_V*
that records T/F for whether the value in column V is an outlier or not.

It is safest to **keep** and **flag** the outliers.

It is a good idea to **check the impact** of the outliers
and your approach to handling the outliers

by comparing the results of your analysis

using the treated outliers
versus removing all the outliers.