Notes: Process Data from Dirty to Clean

Wonnetz Phanthavong

email: wonnetz2@gmail.com.com
linkedin: https://www.linkedin.com/in/wonnetz/

Abstract

- Data Integrity
- Sample Size and Random Sampling
- Confidence Intervals and Margin of Error
- Testing Data
- Clean Data
- Data Cleaning Techniques
- Data Cleaning in both Spreadsheets and Databases
- Verify and report your cleaning results
- Tips for Building a Resume

Week 1: The Importance of Integrity

What is data integrity?

- The accuracy, completeness, consistency and trustworthiness of data throughout it's lifecycle

How can data integrity be compromised?

- Data Replication
- Data Transfer
- Data Manipulation

What to look for if your data aligns with your business objective?

- When there is clean data and good alignment, you can get accurate insights and make conclusions the data supports.
- If there is good alignment but the data needs to be cleaned, clean the data before you perform your analysis.
- If the data only partially aligns with an objective, think about how you could modify the objective, or use data constraints to make sure that the subset of data better aligns with the business objective.

Types of Insufficient Data

- Data from only one source
- Data that keeps updating
- Outdated data
- Geographically limited data

Ways to Address Insufficient Data

- Identify trends with the avilable data
- Wait for more data if time allows
- Talk with your stakeholders and adjust your objectives
- Look for a new dataset

How to determine what Sample Size you should use

- With accordance to the Central Limit Theorem, the smallest representative sample size is given as 30, minimum.

- A confidence level indicates the probability that the same results will be achieved if the analysis were to be taken again.
- In terms of population, this indicates how representative a sample size is.
- The confidence level most commonly used is 95%, but 90% can work in some cases
- A larger sample size can...
  - Increase your confidence level
  - Decrease the margin of error
  - Improve the statistical significance
- sample Size calculators exist, linked, to help you determine your Sample Size.

What is statistical significance?

- Statistical significance is the probability that your analysis is correct. Generally we want a statistical power of .8, or 80% for our analysis to be considered statistically significant.
  ## What is margin of error?
- Margin of error is the maximum amount that the sample results are expected to differ from those of the actual population
- How close are the results of our sample compared to what the results of the overall population would be?

What do you need to calculate the margin of error?

- Population Size
- Sample Size
- Confidence Level

Relationship between Confidence Level and Margin of Error

Youtube Video

Week 2: Sparkling-Clean Data

What is Dirty data?
Data that is incomplete, inconsistent, or irrelevant to the problem at hand.

Types of Dirty Data

- Duplicate Data
- Outdated Data
- Incomplete Data
- Incorrect/Inaccurate Data
- Inconsistent Data

Common Data-Cleaning Pitfalls

- Not checking for spelling errors
- Forgetting to document errors
- Not checking for misfielded values
- Overlooking missing values
- Looking at a subset of data and not the whole picture
- Losing track of the business objective
- Not fixing the source of the error
- Not analyzing the system prior to data cleaning
- Not backing up your data prior to data cleansing
- Not accounting for data cleaning in your deadlines/process

Tips for Cleaning Data

Automating Scientific Data Analysis

Basic Data Cleaning Tools

- Data Validation
- Conditional Formatting
- COUNTIF

- Sorting
- Filtering

Week 3: Using SQL to Clean Data

- Different data cleaning functions in spreadsheets and SQL
- How SQL can be used to clean large data sets
- Apply basic SQL functions for transforming data and cleaning strings

Basic Types of SQL Queries

- SELECT
- INSERT TO
- UPDATE
- CREATE TABLE IF NOT EXISTS
- DROP TABLE IF EXISTS

Cleaning Stringle Variables in SQL

- Using DISTINCT in SELECT
- LENGTH()
- SUBSTR()
- TRIM()

Advanced Data-cleaning Functions in SQL

- CAST()
- CONCAT()
- COALESCE()

Week 4: Verifying and Reporting Results

How can we verify that the data-cleaning effort was well-executed and the resulting data is accurate and reliable?
How do we go about reporting these results?

Take a step back
- Consider the business problem
- Consider the goal
- Consider the data
- Do the numbers make sense?

What can Pivot Tables do?

- Verify Data
- Summarize Data
- Visualize Data

Cleaning Data in SQL

We can use CASE in the SELECT portion of our query to see which entries are clean and which entries still have errors.

The Most Common Problems

- Sources of Errors
  - What is causing these errors? Human? or faulty code?
- Null data
  - Did we use conditional formatting/filters to find all the NULL values?
- Misspelled Words
  - Find/Replace tool
- Mistyped numbers
  - Did we check that the numeric data has been entered in correctly?
- Extra Spaces and Characters
  - TRIM function or SUBSTR function
- Duplicates
  - Remove Duplicates in Spreadsheets or DISTINCT in SQL
- Mismatched Data Types

  - CAST

- Inconsistent Strings
  - Are the strings meaningful and consistent?

- Inconsistent Date Formats
  - Are the year, month, and day consistent?

- Misleading Variable Columns
  - Did we name the columns meaningfully?

- Truncated Data
  - Did we check for truncated or missing data that needs correction?

- Business Logic
  - Does the data make sense in parallel with the business knowledge?

Why should we document changes we've made after cleaning the data?

- Recover cleaning data errors
- Informs other users to the changes we've made
- Determines the quality of the data to be used in analysis

Information to put in a changelog

- Date, file, formula, query, or any other component that changed
- Description of what changed
- Date of the change
- Person who made that change
- Person who approved the change
- Version number
- Reason for the change

Week 5: About the Data-Analyst Hiring Process

Resume Tips

- P: Problem
- A: Action
- R: Result