## **Course content:**

Course 4 – Process Data from Dirty to Clean

1. **Ensuring data integrity.** Data integrity is necessary to ensure a successful analysis. In this part of the course, you will explore methods and steps that analysts take to check data for integrity. This includes knowing what to do when you have an insufficient amount of data. You will also learn about sample size, avoiding sample bias, and using random samples. All of these measures also help to ensure a successful data analysis.
2. **Understanding clean data.** Every data analyst wants clean data to work with when performing an analysis. In this part of the course, you will learn the difference between clean and dirty data. You will practice data cleaning techniques in spreadsheets and other tools.
3. **Cleaning data using SQL.** Knowing a variety of ways to clean data can make an analyst’s job much easier. In this part of the course, you will use SQL to clean data from databases. You will explore how SQL queries and functions can be used to clean and transform your data before an analysis.
4. **Verifying and reporting cleaning results.** Cleaning data is an important step in the data analysis process. In this part of the course, you will verify that data is clean and report data cleaning results. With verified clean data, you will be ready for the next step in the data analysis process.
5. **(Optional) Adding data to your resume.** Creating an effective resume will help you in your data analytics career. In this part of the course, you will learn all about the job application process. Your focus will be on building a resume that highlights your strengths and relevant experience.
6. **Completing the Course Challenge.** At the end of this course, you will be able to apply what you have learned in the Course Challenge. The Course Challenge will ask you questions about the key concepts and then will give you an opportunity to put them into practice as you go through prepared scenarios.

### 

#### **What you will learn:**

* Data integrity and the importance of clean data
* The tools and processes used by data analysts to clean data
* Data-cleaning verification and reports
* Statistics, hypothesis testing, and margin of error
* Resume building and interpretation of job postings (optional)

#### **Skill sets you will build:**

* Connecting business objectives to data analysis
* Identifying clean and dirty data
* Cleaning small datasets using spreadsheet tools
* Cleaning large datasets by writing SQL queries
* Documenting data-cleaning processes

## **Reference: Data constraints and examples**

As you progress in your data journey, you'll come across many types of data constraints (or criteria that determine validity). The table below offers definitions and examples of data constraint terms you might come across.

| **Data constraint** | **Definition** | **Examples** |
| --- | --- | --- |
| **Data type** | Values must be of a certain type: date, number, percentage, Boolean, etc. | If the data type is a date, a single number like 30 would fail the constraint and be invalid |
| **Data range** | Values must fall between predefined maximum and minimum values | If the data range is 10-20, a value of 30 would fail the constraint and be invalid |
| **Mandatory** | Values can’t be left blank or empty | If age is mandatory, that value must be filled in |
| **Unique** | Values can’t have a duplicate | Two people can’t have the same mobile phone number within the same service area |
| **Regular expression (regex) patterns** | Values must match a prescribed pattern | A phone number must match ###-###-#### (no other characters allowed) |
| **Cross-field validation** | Certain conditions for multiple fields must be satisfied | Values are percentages and values from multiple fields must add up to 100% |
| **Primary-key** | (Databases only) value must be unique per column | A database table can’t have two rows with the same primary key value. A primary key is an identifier in a database that references a column in which each value is unique. More information about primary and foreign keys is provided later in the program. |
| **Set-membership** | (Databases only) values for a column must come from a set of discrete values | Value for a column must be set to Yes, No, or Not Applicable |
| **Foreign-key** | (Databases only) values for a column must be unique values coming from a column in another table | In a U.S. taxpayer database, the State column must be a valid state or territory with the set of acceptable values defined in a separate States table |
| **Accuracy** | The degree to which the data conforms to the actual entity being measured or described | If values for zip codes are validated by street location, the accuracy of the data goes up. |
| **Completeness** | The degree to which the data contains all desired components or measures | If data for personal profiles required hair and eye color, and both are collected, the data is complete. |
| **Consistency** | The degree to which the data is repeatable from different points of entry or collection | If a customer has the same address in the sales and repair databases, the data is consistent. |