UNIVERSITY OF SCIENCE

FACULTY OF INFORMATION TECHNOLOGY

PROCESS DATA FROM DIRTY TO CLEAN

COURSERA

Triệu Nhật Minh — 21127112 — 21KHMT2

Instructors

Bùi Duy Đăng
Phạm Trọng Nghĩa
Nguyễn Ngọc Đức

December 15, 2023

Table of Contents

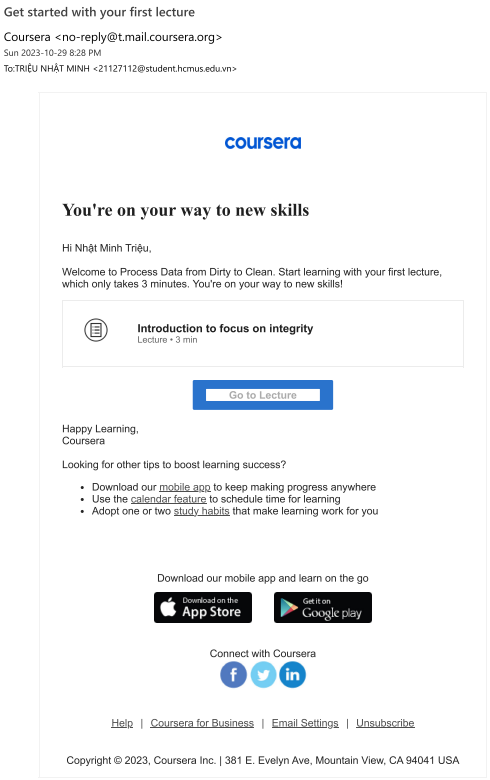
1 Preface

I am deeply appreciative of the chance to be a beneficiary of the 2022 Digital Talent Scholarship, sponsored by NIC. This scholarship has been instrumental in allowing me to delve deeper into my interest in data science and further develop my abilities in this area. A notable aspect of this scholarship is its ongoing nature, allowing me to continually engage with a variety of online courses. I have opted to undertake the Google Data Analytics course.

In order to obtain the Professional Certificate, I am required to complete 8 courses. Each of these courses comprises 4-5 modules, with each module containing 1-2 quizzes. These quizzes present a moderate level of difficulty, necessitating a significant investment of time for completion. Among the courses, Course 4: Process data from dirty to clean, stands out as my favorite due to its practicality. This course has equipped me with numerous valuable skills. I have compiled the main takeaways from this course in this report, with the hope that it will serve as a useful resource for you.

2 Certificate & Enrollment date

Enrollment date: 29th October 2023



Get started with your first lecture

Coursera <no-reply@t.mail.coursera.org>
Sun 2023-10-29 8:28 PM
To:TRIỆU NHẬT MINH <21127112@student.hcmus.edu.vn>

coursera

You're on your way to new skills

Hi Nhật Minh Triệu,

Welcome to Process Data from Dirty to Clean. Start learning with your first lecture, which only takes 3 minutes. You're on your way to new skills!

Introduction to focus on integrity
Lecture • 3 min

Go to Lecture

Happy Learning,
Coursera

Looking for other tips to boost learning success?

- Download our mobile app to keep making progress anywhere
- Use the calendar feature to schedule time for learning
- Adopt one or two study habits that make learning work for you

Download our mobile app and learn on the go

Download on the App Store | Get it on Google play

Connect with Coursera

Help | Coursera for Business | Email Settings | Unsubscribe

Copyright © 2023, Coursera Inc. | 381 E. Evelyn Ave, Mountain View, CA 94041 USA

Completion date of each course: On certificate verification link

3 Course 4: Process Data from Dirty to Clean

3.1 Module 1: The importance of integrity

3.1.1 Data integrity and analytics objectives

Data integrity

Data integrity, which is the extent to which data is valid, complete, consistent, and accurate, is crucial for reliable and meaningful data analysis. Various factors such as different data formats, data replication, data transfer, data manipulation, and human errors can compromise data integrity. However, it can be maintained by employing common standards, verifying data sources, cleaning data, and auditing results. It's important to ensure that data integrity and analytics objectives are well-aligned. This means that the data should adequately support the purpose and goals of the analysis.

Imagine you are a data analyst for a multinational company that operates in different countries. You need to work with data that contains dates, but the dates are not formatted consistently. Depending on the country, the dates might be written as DD/MM/YY, YYYY-MM-DD, or MM/DD/YY. This can cause a lot of confusion and errors in your analysis. For example, you might order extra supplies for the wrong month, or miss an important deadline, or mix up the records of different customers. To avoid these problems, you need to ensure that the data has integrity, which means that it is valid, complete, and clean.

To achieve data integrity, you need to follow some best practices, such as:

- Checking the data type and range of the dates

- Making sure the dates are not missing or duplicated
- Using a standard format for the dates across all sources and systems
- Verifying the accuracy and completeness of the data
- Maintaining the consistency of the data

Well-aligned objectives and data

Clean data + alignment to business objective = accurate conclusions

Impress Me, an online content subscription service, aims to understand the time frame within which users start viewing content after their subscriptions are activated. The data analyst verifies the data's cleanliness and availability, confirming its alignment with the business objective. The only missing piece is the exact duration it takes for each user to start viewing content post-activation.

The data processing steps are illustrated using a user from V&L Consulting. These steps are repeated for each subscribing account and its associated users. The steps include looking up the activation date for the account, identifying a user belonging to the account, finding the user's first content access date, and calculating the time between activation and first content usage.

The analyst could use the VLOOKUP function to look up data in Steps 1, 2, and 3, saving time by avoiding manual lookups. The DATEDIF function could be used in Step 4 to automatically calculate the difference between the dates, providing the number of days between two dates. Both functions are available in Google Sheets and Excel, with the DAYS360 function offering a similar feature in accounting spreadsheets that use a 360-day year.

Pro Tip 1: In the process mentioned above, an analyst could significantly streamline their workflow by utilizing the VLOOKUP function. This function, available in spreadsheet software like Google Sheets, allows the analyst to search for a specific value in a column and return a corresponding piece of information. For instance, in Steps 1, 2, and 3, VLOOKUP could be used to look up data and populate the values in the spreadsheet in Step 4. This function can save a considerable amount of time as it eliminates the need for manual lookup of dates and names.

Pro Tip 2: In Step 4 of the process, the analyst could further optimize their workflow by using the DATEDIF function. This function automatically calculates the difference between the dates in column C and column D, effectively determining the number of days between two dates. For spreadsheets that use a 360-day year (twelve 30-day months), such as those used in accounting, the DAYS360 function serves a similar purpose. Both these functions are available in spreadsheet software like Excel and Google Sheets, and can be a valuable tool for analysts.

Alignment to business objective + additional data cleaning = accurate conclusions

Cloud Gate, a software company, recently conducted a series of free webinars to introduce

their products. The data analyst and the webinar program manager aim to identify companies with five or more attendees for these sessions. The goal is to provide sales managers with this list for potential follow-ups and sales opportunities.

The webinar attendance data includes mandatory fields like the attendee's name and email address, and an optional field for the company name. However, before analyzing the data, the analyst and program manager decide to clean it for accuracy.

Since the company name was not a required field, they plan to infer it from the email address if it's missing. For instance, if the email address is username@google.com, they would fill in 'Google' as the company name. This assumes that attendees using company email addresses attended the webinar for business reasons.

Additionally, they noticed that attendees could enter variations of their names. To ensure accurate counting across multiple webinars, they plan to validate names against unique email addresses. For example, if 'Joe Cox' and 'Joseph Cox' attended two webinars with the same email address, they would be recognized as the same person, ensuring accurate attendee count.

Alignment to business objective + newly discovered variables + constraints = accurate conclusions

A+ Education, an after-school tutoring company, is interested in determining the minimum number of tutoring hours required for students to achieve at least a 10% improvement in their assessment scores. The data analyst believes that the available data, which includes logged tutoring hours and regularly recorded assessment scores, aligns well with this objective.

However, upon closer examination of the data, the analyst identifies additional variables that need to be considered. They notice that while some students had consistent weekly sessions, others had more sporadic schedules, despite having the same total number of tutoring hours. This discrepancy led the analyst to realize that the data might not align as closely with the original business objective as initially thought.

To address this, the analyst decides to add a data constraint to focus solely on students with consistent weekly sessions. This adjustment aims to provide a more accurate understanding of the necessary tutoring duration to achieve a 10% improvement in assessment scores.

3.1.2 Overcoming the challenges of insufficient data

What to do when you find an issue with your data

No data

In real-life scenarios, when faced with the challenge of insufficient data, there are a few possible solutions. One approach is to gather data on a small scale for a preliminary analysis, and then request additional time to complete the analysis after more data has been collected. For instance, if you're surveying employees about their opinions on a new performance and bonus plan, you

could use a sample for an initial analysis, and then ask for another three weeks to gather data from all employees.

If time constraints prevent the collection of more data, another common workaround is to perform the analysis using proxy data from other datasets. For example, if you're analyzing peak travel times for commuters but lack data for a specific city, you could use data from another city of similar size and demographic. This approach allows for the continuation of the analysis while ensuring a reasonable degree of accuracy.

Too little data

In situations where data is lacking, there are several practical solutions that can be implemented. One strategy is to initially gather data on a smaller scale for a preliminary analysis. For instance, if you're conducting a survey on employee feedback regarding a new performance and bonus plan, you could start with a sample group for an initial analysis. Following this, you could request an additional three weeks to collect responses from all employees.

Alternatively, if time constraints prevent further data collection, you could use proxy data from other datasets for your analysis. For example, if you're studying peak commuter travel times but lack data for a specific city, you could use data from another city of similar size and demographic. This method allows for the continuation of the analysis while maintaining a reasonable degree of accuracy.

Wrong data, including data with errors

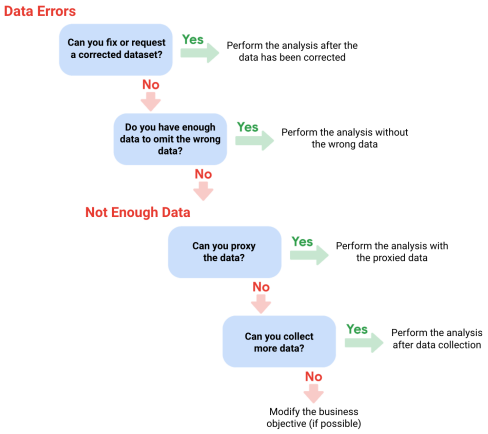
In real-life scenarios, when faced with the challenge of having incorrect data, there are several practical solutions. If the wrong data was received due to misunderstood requirements, it's important to re-communicate the requirements. For instance, if data for male voters was received when the requirement was for female voters, the needs should be restated.

Errors in the data should be identified and, if possible, corrected at the source by looking for a pattern in the errors. For example, if data is in a spreadsheet and a conditional statement or boolean is causing incorrect calculations, the conditional statement should be changed instead of just fixing the calculated values.

If data errors cannot be corrected personally, the wrong data can be ignored and the analysis can proceed if the sample size is still large enough and ignoring the data won't cause systematic bias. For instance, if a dataset was translated from a different language and some of the translations don't make sense, the data with bad translation can be ignored and the analysis can continue with the other data.

Calculate sample size

One of the challenges of data analysis is determining the appropriate sample size for your project. A sample size is the number of observations or units from a larger population that you



use to estimate some characteristics of the population. There are some general guidelines to follow when choosing a sample size, such as:

- Avoid using a sample size less than 30, as this may not be representative of the population and may not follow the normal distribution. This is based on the Central Limit Theorem, which states that as sample size increases, the sample mean approaches the population mean.
- Consider the confidence level you want to achieve, which is how confident you are that your sample results are close to the true population results. The most common confidence level is 95%, meaning you are 95% sure that your sample results are within a certain range of the population results. If you want a higher confidence level, you will need a larger sample size.
- Think about the margin of error you can tolerate, which is how much your sample results may differ from the population results. The smaller the margin of error, the more precise your sample results are. If you want a smaller margin of error, you will need a larger sample size.
- Assess the statistical significance you want to achieve, which is how likely your sample results are not due to chance. The higher the statistical significance, the more confident you are that your sample results reflect a real difference or relationship in the population. If you want a higher statistical significance, you will need a larger sample size.

The size of your sample can vary significantly depending on the specific business problem you are trying to solve. For instance, if you were to conduct a survey in a city with a population of 200,000 and managed to get 180,000 responses, that would constitute a large sample size. However, in practical terms, you might be wondering what an acceptable, smaller sample size might look like.

If the people surveyed represented every district in the city, a sample size of 200 might be sufficient. However, the adequacy of this sample size really depends on the stakes of your business

problem. For example, a sample size of 200 might be large enough if you're trying to gauge how residents feel about a new library. But if you're trying to determine how residents would vote to fund the library, a sample size of 200 might not be large enough. In this case, you might be more willing to accept a larger margin of error when surveying residents' feelings about the new library, as opposed to surveying residents about how they would vote to fund it.

It's also important to consider the cost of obtaining a larger sample size. While larger sample sizes can yield more accurate results, they also come with a higher cost. For instance, someone trying to understand consumer preferences for a new line of products might not need as large a sample size as someone trying to understand the effects of a new drug. In the case of drug safety, the benefits of more accurate results might outweigh the cost of using a larger sample size. However, for consumer preferences, a smaller sample size at a lower cost could provide sufficiently accurate results.

Knowing the basics of sample size determination can help you make the right choices. If you come across a sample size that seems too small, you can always raise concerns. Tools like a sample size calculator can be very helpful in this regard. These calculators allow you to enter a desired confidence level and margin of error for a given population size, and then calculate the sample size needed to statistically achieve those results. You can refer to resources such as the "Determine the Best Sample Size" video or the "Sample Size Calculator" reading for additional information.

3.1.3 Testing your data

What to do when there is no data

Proxy data examples

Proxy data can be a valuable tool when the necessary data to support a business objective isn't readily available. For instance, an auto dealership that has just launched a new car model might not want to wait until the end of the month for sales data. In this case, the number of clicks on the car specifications on the dealership's website can serve as a proxy for potential sales. Similarly, a supplier of a new plant-based meat product can use the sales data of a tofu-based turkey substitute to estimate future demand. Lastly, the Chamber of Commerce, wanting to assess the impact of a tourism campaign, can use historical data for airline bookings following a similar campaign as a proxy. These examples illustrate how proxy data can provide immediate insights in various business scenarios.

Open (public) datasets

This section discusses the utilization of open or public datasets for the purpose of data analysis. An illustration is provided where a medical clinic leverages an open dataset from a vaccine trial to approximate the count of contraindications for a nasal vaccine. The text further intro-

duces Kaggle, a platform renowned for hosting a variety of dataset formats including CSV, JSON, SQLite, and BigQuery. Users are encouraged to consult the Kaggle documentation and independently explore the datasets available. A cautionary note is also included, urging users to be vigilant of duplicate data and Null values within open datasets, and to comprehend the usage of Null prior to initiating data analysis.

Sample size calculator

How to use a sample size calculator

To utilize a sample size calculator effectively, it's essential to have certain parameters determined beforehand. These include the size of the population you're studying, the desired confidence level, and an acceptable margin of error. Once you have these details, you can input them into a sample size calculator. There are several such calculators available online, including those offered by platforms like SurveyMonkey and Raosoft.

What to do with the results

The outcomes from a sample size calculator indicate the minimum number of individuals or items you need to examine to obtain dependable results. You also need to take into account the number of people who will actually respond to your survey, if you are using one. For example, to receive 100 responses with a 10% response rate, you need to distribute your survey to 1,000 people. You can hone your skills using the sample size calculators and revisit the terms in this reading if necessary.

3.1.4 Consider the margin of error

All about margin of error

Margin of error is the maximum amount that the sample results are expected to differ from those of the actual population. More technically, the margin of error defines a range of values below and above the average result for the sample. The average result for the entire population is expected to be within that range. We can better understand margin of error by using some examples below.

Want to calculate the margin of error

To effectively use a margin of error calculator, you need to understand a few key terms. The confidence level is a percentage that shows how likely your sample accurately represents the larger population. The population is the total number from which you draw your sample, and the sample is a portion of the population that represents the whole. The margin of error is the maximum amount that the sample results may differ from the actual population.

Typically, a confidence level of 90% or 95% is used. However, some industries may require a stricter confidence level, such as 99%, which is common in the pharmaceutical industry.

Once you've determined your population size, sample size, and confidence level, you can input this information into a margin of error calculator. There are several available online, including those offered by Good Calculators and CheckMarket.

Key takeaway

Margin of error is used to determine how close your sample's result is to what the result would likely have been if you could have surveyed or tested the entire population. Margin of error helps you understand and interpret survey or test results in real-life. Calculating the margin of error is particularly helpful when you are given the data to analyze. After using a calculator to calculate the margin of error, you will know how much the sample results might differ from the results of the entire population.

3.1.5 Module 1: Module challenge

Back Module 1 challenge
Graded Quiz • 40 min

English Due Nov 5, 11:59 PM +07

Congratulations! You passed!

Grade received 100% Latest Submission Grade 100% To pass 80% or higher

Go to next item

1. Fill in the blank: As a data analyst, you need to verify that your data is _____ to ensure your analysis and conclusions are accurate. 1 / 1 point

- manipulated and replicated
- manipulated and valid
- private and valid
- complete and valid

Correct

2. What is one potential problem associated with data manipulation that analysts must be aware of? 1 / 1 point

- Data manipulation can help organize a dataset.
- Data manipulation can separate a dataset among different locations.
- Data manipulation can introduce errors.
- Data manipulation can make a dataset easier to read.

Correct

3. A data analyst is given a dataset for analysis. It includes data only about the total population of every country in the previous 20 years. Based on the available data, an analyst would have the full picture and be able to determine the reasons behind a certain country's population increase from 2016 to 2017. 1 / 1 point

- True
- False

Correct

4. A data analyst is given a dataset for analysis. To use the template for this dataset, click the link below and select "Use Template." 1 / 1 point

Link to template: June 2014 Invoices

OR

If you don't have a Google account, download the CSV file directly from the attachment below.

June 2014 Invoices - Sheet1
CSV File

Which of the following has duplicate data?

- Data for Valando on 1/1/2014
- Data for Valando on 2/18/2014
- Data for Symteco on 2/21/2014
- Data for Symteco on 5/20/2014

Correct

5. A data analyst at a nonprofit organization is working with a dataset about a summer fundraiser. Although they have a lot of useful data by the end of June, they recognize that the data is insufficient. So, they decide to wait until the end of the season to begin working with the dataset. Which type of insufficient data does this example describe? 1 / 1 point

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

Correct

6. A clothing manufacturer wants to learn more about why their consumers have purchased the brand's products. How should this manufacturer conduct their survey? 1 / 1 point

- Send the survey to their least frequent customers
- Send the survey to random people who buy clothes
- Send the survey to a representative sample of their customers
- Send the survey to customers who have purchased more than one product

Correct

7. A car dealership gathers data about their entire customer population. They decide to conduct a survey to understand why their customers chose their dealership. They send out an email to all customers who have purchased more than two vehicles in the past five years. What does this scenario describe? 1 / 1 point

- Geographically limited sampling
- Unbiased sampling
- Random sampling
- Sampling bias

Correct

8. Sometimes during analysis, an analyst discovers that it's necessary to adjust the business objective. When this happens, the analyst should take the initiative to do so without involving others in order to be respectful of their time. 1 / 1 point

- True
- False

Correct

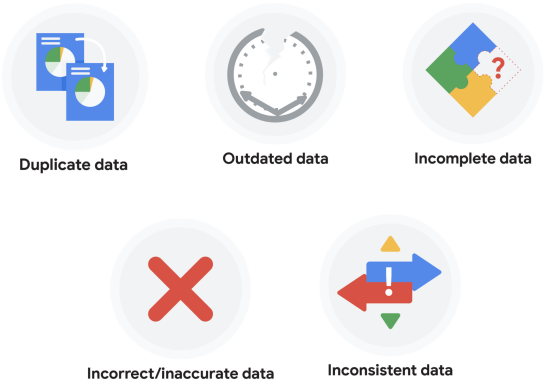
3.2 Module 2: Sparkling-clean data

3.2.1 Data cleaning is a must

What is dirty data?

Types of dirty data Data quality issues can take various forms, each with its own causes and potential harm to businesses.

- Duplicate data which refers to any data record that appears more than once, can be caused by manual data entry, batch data imports, or data migration. This can lead to skewed metrics or analyses, inflated or inaccurate counts or predictions, and confusion during data retrieval.
- Outdated data is any data that is old and should be replaced with newer, more accurate information. This can occur when people change roles or companies, or when software and systems become obsolete. The result can be inaccurate insights, decision-making, and analytics.
- Incomplete data refers to any data that is missing important fields, often due to improper data collection or incorrect data entry. This can decrease productivity, lead to inaccurate insights, or prevent the completion of essential services.
- Incorrect or inaccurate data is any data that is complete but inaccurate, which can be due to human error during data input, fake information, or mock data. This can lead to inaccurate insights or decision-making based on bad information, resulting in revenue loss.
- Inconsistent data is any data that uses different formats to represent the same thing. This can occur when data is stored incorrectly or errors are inserted during data transfer. The result can be contradictory data points leading to confusion or an inability to classify or segment customers.



Business impact of dirty data

For additional information on the business consequences of unclean data, type "dirty data" into the search bar of your preferred web browser. This will yield a multitude of articles on the subject. The following are some impacts mentioned for specific sectors from a prior search:

- Banking:Inaccuracies cost companies between 15% and 25% of revenue (source)
- Digital commerce: Up to 25% of B2B database contacts contain inaccuracies (source)
- Marketing and sales: 99% of companies are actively tackling data quality in some way (source)
- Healthcare: Duplicate records can be 10% and even up to 20% of a hospital's electronic health records (source)

3.2.2 Begin cleaning data

Data cleaning tools and techniques

The topic has taught me that maintaining clean data is essential for preserving data integrity and facilitating reliable decision-making. There are numerous tools and techniques in spreadsheets that aid in data cleaning. It's advisable to make a copy of the dataset before initiating the cleaning process for future reference. Duplicates, irrelevant data, extra spaces, blanks, typos, inconsistencies, and formatting issues are common problems in datasets.

To tackle these, spreadsheet tools can be used to automatically identify and eliminate duplicates. Typos require manual correction, while clear format tools assist in maintaining a visually consistent appearance. It's crucial to remove irrelevant data that doesn't contribute to the problem under investigation. Extra spaces and blanks can lead to unexpected results when sorting, filtering, or searching through data, hence should be removed.

Inconsistencies in capitalization, incorrect punctuation, and other typos can cause significant issues, especially when handling sensitive data like emails. Spreadsheet tools such as spellcheck, autocorrect, and conditional formatting can be employed to rectify these issues.

Lastly, when dealing with data from various sources, removing formatting is important as each database has its unique formatting. Using a "clear formats" tool can help achieve a clean and visually consistent appearance for your spreadsheets. This knowledge is fundamental for anyone working with data as it ensures the reliability of the analysis and decision-making processes.

Cleaning data from multiple sources

In this lecture, I've gained insights into the complexities and challenges involved in cleaning and merging multiple datasets, a common task for data analysts. The speaker provided a real-world example of a potential merger between the International Logistics Association and the Global Logistics Association, emphasizing the need for data merging when organizations unite.

The process of data merging involves combining two or more datasets into a single dataset. The challenges arise from the inherent inconsistencies and misalignments when merging datasets with different structures. The example illustrated issues such as inconsistent address formats, varying member ID formats, and differences in membership types between the two associations.

I've learned that data analysts must address these challenges by standardizing information, removing duplicates, and harmonizing terminology before the datasets can be effectively merged. The importance of compatibility, defined as how well datasets can work together, was highlighted. Key questions were presented, such as ensuring all necessary data is available, confirming the existence of required data within the datasets, and assessing whether the datasets are clean and adhere to the same standards.

The lecture also touched on the importance of examining fields that are regularly repeated, handling missing values, and considering the recency of data updates. The choice of tools for data cleaning and merging, including spreadsheet tools, SQL queries, and programming languages like R, was briefly discussed, with a promise of further exploration in upcoming sessions. Overall, I've gained a foundational understanding of the considerations and steps involved in effectively merging datasets for analysis.

Common data-cleaning pitfalls

- **Not checking for spelling errors**: Typographical or input errors often result in misspellings. Frequently, incorrect spellings or common grammatical mistakes can be identified, but it becomes more challenging with unique elements like names or addresses. For instance, while dealing with a customer data table in a spreadsheet, you may encounter a customer named "John" whose name has been mistakenly entered as "Jon" in certain instances. The spreadsheet's spellcheck is unlikely to highlight this, so if you don't meticulously check for spelling errors, your analysis may contain inaccuracies.

- **Forgetting to document errors**: Keeping a record of your mistakes can significantly save time, as it aids in preventing future errors by demonstrating how you rectified them. For instance, you may encounter a mistake in a formula in your spreadsheet. You realize that the dates in one of your columns are not formatted properly. By documenting this correction, you can refer to it the next time your formula malfunctions, giving you a jumpstart on problem-solving. Error documentation also assists in monitoring alterations in your work, enabling you to revert changes if a solution was ineffective.

- **Not checking for misfielded values**: Misfielded values occur when entries are placed in incorrect fields. These entries might still adhere to the correct format, making them more challenging to identify without meticulous attention. For instance, consider a dataset with separate columns for cities and countries. Since these data types are similar, they can easily be confused. If you were searching for all instances of Spain in the country column,

but Spain was erroneously inputted in the city column, you would overlook crucial data points. Ensuring the accuracy of your data entry is vital for comprehensive and precise analysis.

- **Overlooking missing values**: Absent values in your dataset can lead to mistakes and yield incorrect results. For instance, if you were attempting to calculate the total sales from the previous three months, but a week's worth of transactions were not included, your computations would be off. It's a good practice to strive for data cleanliness by ensuring its completeness and uniformity.

- **Only looking at a subset of the data**: It's crucial to consider all pertinent data during the cleaning process. This ensures a comprehensive understanding of the narrative the data presents and a thorough check for potential errors. For instance, if you're dealing with bird migration data from various sources and only clean one source, you might overlook repeated data. This oversight can lead to issues in your subsequent analysis. To prevent common errors like duplicates, every data field deserves equal scrutiny.

- **Losing track of business objectives**: While cleaning data, you may stumble upon intriguing findings about your dataset. However, these discoveries shouldn't divert you from your primary task. For example, while working with weather data to determine your city's average rainy days, you might spot interesting snowfall patterns. While this is fascinating, it's not pertinent to your current query. Curiosity is commendable, but it shouldn't distract you from your main task.

- **Not fixing the source of the error**: Rectifying the error is vital, but if the error is a symptom of a larger issue, you need to identify the root cause. Otherwise, you'll find yourself repeatedly fixing the same error. Suppose you have a team spreadsheet tracking everyone's progress, and the table keeps breaking due to inconsistent entries. You could continually fix these issues individually, or you could streamline data entry to ensure everyone is aligned. Tackling the root cause of your data errors will save you considerable time in the long run.

- **Not analyzing the system prior to data cleaning**: To clean our data and prevent future errors, we need to understand the origin of the dirty data. Just like an auto mechanic would identify the problem before starting repairs, the same applies to data. Initially, you determine the error sources, which could be data entry errors, lack of spell check, absence of formats, or duplicates. Once you comprehend where the bad data originates, you can manage it and maintain clean data.

- **Not backing up your data prior to data cleaning**: It's always beneficial to be proactive and back up your data before initiating data clean-up. If your program crashes or if your modifications create a dataset issue, you can always revert to the saved version. The simple

act of backing up your data can save you hours of work and, most importantly, prevent headaches.

- **Not accounting for data cleaning in your deadlines/process**: Good things, including data cleaning, require time. It's essential to remember this when planning your process and setting deadlines. Allocating time for data cleaning allows for more accurate ETA estimates for stakeholders and helps you know when to request an adjusted ETA.

Data cleansing is key for accurate analysis. Common pitfalls include typographical errors, misplaced or missing values, not examining all data, losing focus of business goals, not rectifying error sources, not assessing the system before cleaning, not backing up data, and not including data cleaning in your timelines. Avoiding these can ensure data integrity, leading to improved business outcomes.

3.2.3 Cleaning data in spreadsheets

Data-cleaning features in spreadsheets

In this lecture, the instructor provided insights into various tools and techniques for cleaning data in spreadsheet applications. The discussion began with an exploration of conditional formatting, a valuable tool that alters cell appearances based on specified conditions. This feature serves as a visual aid, making it easier for data analysts to identify and understand information in large datasets, particularly when certain data points do not meet specified conditions.

A practical demonstration in a logistics association spreadsheet showcased the application of conditional formatting to highlight blank cells, effectively signaling missing information for further attention and addition to the spreadsheet.

The lecture then delved into the tool for removing duplicates, emphasizing the importance of creating a copy of the dataset before utilizing this feature. A step-by-step guide was provided on using the "Remove duplicates" tool, exemplified by the removal of a duplicated entry in the association member list.

Consistency in formatting, especially for dates, was discussed as another crucial aspect of data cleaning. The instructor demonstrated how to make date formats consistent, ensuring clarity for analytical purposes.

The lecture also introduced the concept of text strings and substrings, paving the way for an explanation of the "Split" tool. This tool proves useful when there is a need to divide a text string around a specified delimiter and allocate each fragment into new and separate cells. The example of splitting professional certifications in a column, separated by commas, illustrated its practical application.

Additionally, the lecture touched on the use of the "Split text to columns" tool to address

instances of numbers stored as text. A cosmetics maker's spreadsheet was used as an example to demonstrate how this tool could resolve errors caused by text-formatted numbers.

The upcoming discussion on the CONCATENATE function was briefly introduced as a tool that performs the opposite action, joining multiple text strings into a single string.

Overall, the instructor conveyed a comprehensive understanding of several common spreadsheet tools used in data cleaning, emphasizing their significance in saving time, ensuring accuracy, and facilitating efficient data analysis in the field of data analytics.

Althought I haved learned these techniques in course "Introduction to information technology" at school, I still find them very useful and practical. As a practical example, I have used the "Remove duplicates" tool to remove duplicate rows in a spreadsheet containing information about participants in my club's event.

Optimize the data cleaning process

In this lecture, the instructor explored spreadsheet functions and their importance in enhancing data cleaning processes to ensure data integrity. Functions, defined as specific instructions that perform calculations using spreadsheet data, were presented as potent tools in data analysis.

The lecture kicked off by emphasizing the COUNTIF function, which counts the number of cells that match a specified value within a given range. Its practical use was demonstrated in a professional association spreadsheet to check for anomalies in membership prices, such as negative numbers or values deviating significantly from expected ranges.

Next, the LEN function was introduced, which counts the characters in a text string, providing insight into its length. This function was used in the association spreadsheet to confirm the correct length of six-digit member identification codes.

Conditional formatting was revisited as an efficient tool to identify instances where the length of text strings did not meet expected criteria. An example with member identification codes showed how the LEN function and conditional formatting could simplify the data cleaning process.

The lecture proceeded with the introduction of LEFT and RIGHT functions, illustrating their use in extracting specific character sets from the left or right side of a text string. A cosmetics maker's spreadsheet was used to demonstrate how these functions can isolate numeric codes and text identifiers.

The MID function was then explained as a tool for extracting a segment from the middle of a text string. An example involved extracting state abbreviations from client codes in a cosmetics company's client list.

The CONCATENATE function was introduced as a method that combines two or more text strings. The instructor showed how to use CONCATENATE to reassemble left and right text strings into complete product codes in a practical example.

The final topic discussed was the TRIM function, highlighting its usefulness in eliminating leading, trailing, and repeated spaces in data. An example with client names in the cosmetics maker's spreadsheet showed how TRIM could improve data cleanliness.

The thorough overview of these functions offered valuable insights into their use for data cleaning, providing practical examples and hands-on demonstrations. The instructor, also conveyed the information in an understandable manner, urging viewers to practice these data cleaning steps and incorporate them into their analytical workflows.

Workflow automation

What can be automated

Automation indeed seems fantastic, but despite its convenience, there are still aspects of the job that cannot be automated. For instance, communication with your team and stakeholders cannot be automated as it is crucial for understanding their needs while you work on tasks. There is simply no substitute for human interaction. Similarly, presenting your findings, a significant part of a data analyst's job, cannot be automated. Making data comprehensible and accessible to stakeholders and creating data visualizations requires a human touch, much like communication.

However, the preparation and cleaning of data can be partially automated. Specific processes, such as using a programming script to automatically detect missing values, can be set up to automate some tasks. Data exploration can also be partially automated. While visualizing data is often the best way to understand it, there are numerous tools available that can automate the visualization process. These tools can expedite the process of visualizing and understanding the data, but the exploration itself still requires a data analyst.

Data modeling, a complex process involving various factors, can be fully automated. There are tools available that can automate the different stages of data modeling, making this aspect of the job significantly more efficient.

One of the most effective strategies to optimize your data cleaning is to clean the data at its source. This not only benefits your entire team but also eliminates the need for repetitive processes. For instance, you could develop a programming script that calculates the word count in each spreadsheet file located in a specific directory. Utilizing tools that operate directly where your data resides means you don't have to redo your cleaning procedures, thereby conserving time and effort for you and your team.

Even more data-cleaning techniques

The instructor gave a lecture on data mapping, a vital concept for data cleaning in analysis. Data mapping aims to match fields from different databases, which is important for data migration, integration, and management. The instructor stressed the need to know how data changes and moves between systems for analysis.

The instructor outlined the steps of data mapping, starting with identifying the data to be moved, such as tables and fields. The next step was to define the format of the data at the destination. The example of combining data from two logistics associations showed how to do data mapping, focusing on member IDs and their format, such as numbers or emails.

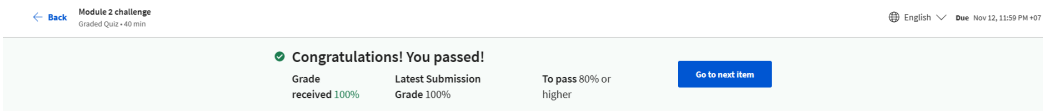
The instructor explained that data mapping can be simple or complex, depending on the schema and the keys. For complex projects, data mapping software tools were mentioned, which can automate data analysis, cleaning, matching, inspection, and validation.

The instructor also discussed the factors to consider when choosing data mapping software, such as support for different file types like Excel, SQL, and Tableau. The importance of consistent naming conventions for data compatibility was emphasized.

The instructor then showed how to map data manually, which involved deciding the content of each section and making the data consistent by transforming it to a common format. The CONCATENATE function was used again as a helpful tool for consistency, shown by an example of merging addresses with different formats.

The lecture ended with the testing phase of data mapping, which required checking a sample of data to ensure its cleanliness and proper formatting. The instructor mentioned familiar data cleaning tools like data validation, conditional formatting, COUNTIF, sorting, and filtering for testing. The instructor highlighted the significance of data mapping in avoiding errors during data merging, which could affect the data quality of the organization.

3.2.4 Module 2: Module challenge



1. Fill in the blank: Conditional formatting is a spreadsheet tool that changes how _____ appear when values meet a specific condition. 1 / 1 point

- filters
- charts
- **cells**
- queries

Correct

2. A data analyst is working on a spreadsheet in which one of the columns contains name data. This data is formatted as *lastname_firstname*. The analyst splits this data at the underscore so that each piece—*firstname* and *lastname*—are contained in their own columns. 1 / 1 point

In this context, what is the underscore acting as?

- Substring
- Partition
- MID function
- **Delimiter**

Correct

3. Fill in the blank: When describing a SUM function, the _____ is =SUM(value 1 through value 2). 1 / 1 point

- structure
- script
- standard
- **syntax**

Correct

4. In a spreadsheet, what is the correct function for extracting the first two characters of the string located in cell A7? 1 / 1 point

- **=LEFT(A7,2)**
- =RIGHT(2,A7)
- =RIGHT(A7,2)
- =LEFT(2,A7)

Correct

5. A data analyst in a human resources department is working with the following selection of a spreadsheet: 1 / 1 point

N/A	A	B	C	D
1	**Year Hired**	**Last 4 of SS#**	**Department**	**Employee ID**
2	2019	1192	Marketing	
3	2014	2683	Operations	
4	2020	1939	Strategy	
5	2009	3208	Graphics	

They want to create employee identification numbers (IDs) in column D. The IDs should include the year hired plus the last four digits of the employee's Social Security Number (SS#). What function will create the ID 20093208 for the employee in row 5?

- **=CONCATENATE(A5,B5)**
- =CONCATENATE(A5|B5)
- =CONCATENATE(A5*B5)
- =CONCATENATE(A5+B5)

Correct

6. An analyst is cleaning a new dataset containing 500 rows. They want to make sure the data contained from cell B2 through cell B300 does not contain a number greater than 50. Choose the statement that includes the correct syntax for this COUNTIF function? 1 / 1 point

- **=COUNTIF(B2:B300,">50")**
- =COUNTIF(B2:B300,"<50")
- =COUNTIF(B2:B300,>50)
- =COUNTIF(B2:B300,<=50)

Correct
In this case, the analyst can use the inequality ">50" to count all the cells containing the value greater than 50 that exist between the cell range B2 through B300. If the return value of the COUNTIF function is greater than the value 0, the analyst can then take the necessary steps to locate and clean the unwanted cell or cells.

7. Fill in the blank: The function _____ is used to return information in a column that contains a specified value. 1 / 1 point

- VALUE
- FIND
- **VLOOKUP**
- MATCH

Correct

8. A data analyst is adding new data to the dataset. This data comes from a different source than the current data. What should the data analyst do to ensure the compatibility of the current and new data? 1 / 1 point

- **Map the data**
- Apply a data structure
- Use a data visualization
- Spot check for null values

Correct

3.3 Module 3: Data cleaning in SQL

3.3.1 Using SQL to clean data

Understanding SQL capabilities

In this video, I learned about SQL, a structured query language that data analysts use to work with large datasets in relational databases. I learned that SQL was developed in the 1970s by IBM computer scientists and became the standard language for relational database communication. I also learned that SQL can handle huge amounts of data, such as trillions of rows, much faster than other tools like spreadsheets. I heard a personal story from the instructor about how he taught himself SQL in a week for a job application and encouraged me to do the same. Finally, I learned that some of the tools we learned in spreadsheets can also be applied to working in SQL.

Using SQL as a junior data analyst

Spreadsheets and SQL databases are both useful tools for data analysis, but they have different strengths and weaknesses.

Spreadsheets are good for working with smaller data sets that can be entered manually. They allow you to create graphs and visualizations within the same program, and they have built-in functions that can help you with data cleaning and formatting. However, spreadsheets are not ideal for working with large data sets that require accessing multiple tables across a database. They can also be slow and inefficient when performing complex calculations. Spreadsheets are best suited for individual projects that do not require collaboration or tracking of queries.

SQL databases are great for working with large data sets that can be accessed from different tables using queries. They have fast and powerful functionality that can handle complex calculations and data manipulation. They can also prepare data for further analysis in other software, such as Tableau or R. However, SQL databases do not allow you to enter data manually or create graphs and visualizations within the same program. They also do not have built-in functions for spell check or other useful features. SQL databases are best suited for collaborative projects that require tracking of queries and working with multiple sources of data.

3.3.2 Learn SQL queries

Cleaning string variables using SQL

While exploring spreadsheets and SQL, we've identified both their unique characteristics and shared features. Despite their differences, there are universal tools that can be used in both to achieve similar results. The techniques learned in spreadsheet data cleaning, such as arithmetic operations, formulas, and data joining, can also be applied in SQL, allowing us to handle more complex tasks.

For example, SQL becomes essential when managing extensive health data in a hospital environment. It provides an efficient way to access and process various data sources, including demographic information, insurance details, public health data, and user-generated data. Given the potential for millions of rows and multiple related tables stored in different formats, SQL is indispensable for managing such intricate datasets.

Unlike manual data entry in spreadsheets, SQL allows us to pull information from different parts of a database. When looking for specific insights, like the count of patients with a certain diagnosis on a specific day, SQL queries using COUNT and WHERE can provide results similar to spreadsheet functions but on a much larger and more complex dataset.

However, it's important to understand that spreadsheets and SQL are fundamentally different. Spreadsheets, created with programs like Excel or Google Sheets, are designed for executing built-in functions. In contrast, SQL is a language designed for interacting with database programs like Oracle MySQL or Microsoft SQL Server. Their primary differences lie in their use cases.

Data analysts often use spreadsheets for data cleaning and analysis when dealing with smaller datasets. But for larger datasets, especially those with over a million rows or spread across multiple database files, SQL is a more efficient, faster, and repeatable choice. Unlike spreadsheets, SQL can automatically access and use data from different parts of a database.

Due to these differences, spreadsheets and SQL serve different functions. Spreadsheets are suitable for smaller datasets and individual work, offering built-in features like spell check. On the other hand, SQL is excellent for managing large datasets, even up to trillions of rows, making it ideal for collaborative work within larger teams accessing data stored across a database. Additionally, SQL's long-standing status as the standard language for database communication increases its compatibility with various database programs and makes it easier to track changes in queries when working collaboratively.

SQL is the main programming language in course "Introduction to database" which I have learned at school, I find this section is quite easy and familiar to me. However, I still find it useful to review the basic concepts of SQL and its applications in data cleaning which the instructor at school did not mention.

3.3.3 Transforming data

Advanced data-cleaning functions (part 1 and 2)

In this learning session, the significance of the CAST function in SQL was emphasized, particularly for data cleaning tasks where data may be imported with incorrect data types. An example was given involving transaction data from Lauren's Furniture Store, where the purchase_price column was incorrectly identified as a string rather than a float.

The tutorial demonstrated how to construct a SQL query to sort purchases by purchase_price

in descending order. It pointed out the problem of incorrect sorting when the database treats numeric values as text strings, resulting in unexpected outcomes. The CAST function was introduced as a solution, showing how it can be used to typecast the purchase_price column to FLOAT64, enabling SQL to correctly identify it as a numerical data type.

The session also provided insights into the concept of typecasting, which is the process of converting data from one type to another. This understanding is not limited to numeric values, as the CAST function can also be used to convert strings into other data types like date and time.

The importance of early data type conversion was stressed, especially when dealing with data from diverse sources. As a data analyst, it is essential to ensure that data is recognizable and usable in the database for accurate and efficient analysis. The CAST function was highlighted as an important tool in the data analyst's toolkit for data cleaning and preparation.

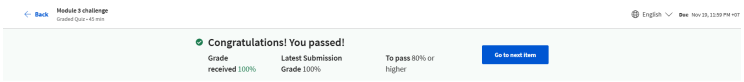
I've gained a deeper understanding of advanced SQL functions, specifically focusing on CAST, CONCAT, and COALESCE.

1. CAST Function
    - Recap: The CAST function was revisited, emphasizing its role in typecasting text strings into different data types, such as converting datetime to date for cleaner results.
    - New Insight: CAST is a versatile tool for cleaning and sorting data, ensuring accurate analysis by addressing issues related to data types.
2. CONCAT Function
    - Introduction: The CONCAT function was introduced as a means to concatenate strings, creating unique keys. It was demonstrated using an example where product colors needed to be distinguished by generating a unique key for analysis.
    - Application: CONCAT is useful for combining string values, helping in scenarios where unique identifiers are required, such as distinguishing product variations.
3. COALESCE Function
    - Introduction: COALESCE was introduced as a function to return non-null values in a list. It is particularly helpful when dealing with optional fields or missing data.
    - Application: COALESCE was demonstrated in creating a list of product names, with a fallback to product codes if the names are null. This ensures more meaningful and readable results.

In conclusion, the 2 videos not only reinforced my understanding of these advanced SQL functions but also provided practical examples that I can apply in my own data analysis work. The

importance of data cleanliness and the role of SQL in handling large datasets were reiterated, setting the stage for the next steps in the analysis process. The video concluded with encouragement to practice and apply these concepts to solidify the learning.

3.3.4 Module 3: Module challenge



1. Fill in the blank: Data analysts usually use _____ to deal with very large datasets. 1 / 1 point

- SQL
- word processors
- spreadsheets
- CSV

Correct

2. What are some of the benefits of using SQL for analysis? Select all that apply. 1 / 1 point

- [x] SQL tracks changes across a team.

Correct

- [x] SQL interacts with database programs.

Correct

- [ ] SQL has better user management than spreadsheets.
- [x] SQL can pull information from different database sources.

Correct

3. A data analyst creates a database to store information on the company's customer data. When completing the initial import the analyst notices that they forgot to add a few customers into the table. What command can the analyst use to add these missed customers? 1 / 1 point

- DROP
- INSERT INTO
- APPEND
- ADD

Correct

4. You are working with a database table named invoice that contains invoice data. The table includes columns for *customer_id* and *total*. You want to remove duplicate customers and identify which unique customers have a total greater than 5. 1 / 1 point

You write the SQL query below. Add a DISTINCT clause that will remove duplicate entries from the *customer_id* column.

NOTE: The three dots (...) indicate where to add the clause.

```
SELECT DISTINCT*
FROM
invoice
WHERE total > 5
```

Run Reset

What customer_id number is located in row 57

NOTE: The query index starts at 1 not 0.

- 57
- 4
- 52
- 14

Correct
The clause **DISTINCT customer_id** will remove duplicate entries from the *customer_id* column. The complete query is **SELECT DISTINCT customer_id FROM invoice WHERE total > 5**.

5. You are working with a database table named customer that contains customer data. The table includes columns about customer location such as *city, state, country* and *postal_code*. The state names are abbreviated. You want to check for city names that are greater than 5 characters long. 1 / 1 point

You write the SQL query below. Add a LENGTH function that will return any city names that are greater than 5 characters long.

NOTE: The three dots (...) indicate where to add the clause.

```
SELECT
*
FROM
customer
WHERE length(city) > 5
```

Run Reset

What is the the last name of the customer in row 1 of your query result?

NOTE: The query index starts at 1 not 0.

- Gonçalves
- Tremblay
- Philips
- Almeida

Correct
The function LENGTH(city) > 5 will return any state names that are greater than 2 characters long. The complete query is SELECT * FROM customer WHERE LENGTH(city) > 5. The LENGTH function counts the number of characters a string contains. The country Ireland is in row 1 of your query result.

6. In SQL databases, what data type refers to a number that does not contain a decimal? 1 / 1 point

- Integer
- Boolean
- String
- Float

Correct

7. A data analyst notices their Boolean column is incorrectly storing True/False values as strings. What function can the analyst use to convert the data type from a string to Boolean? 1 / 1 point

- LENGTH
- CAST
- TRIM
- SUBSTR

Correct

8. After joining multiple tables you find your data contains a significant amount of null values. What function can you use to return only the non-null values in a list? 1 / 1 point

- TRIM
- CONCAT
- CAST
- COALESCE

Correct

9. You are working with a database table that contains employee data. The table includes columns about employee location such as *city, state, country,* and *postal_code*. You use the SUBSTR function to retrieve the first 3 characters of each last_name, and use the AS command to store the result in a new column called *new_last_name*. 1 / 1 point

You write the SQL query below. Add a statement to your SQL query that will retrieve the first 3 characters of each *last_name* and store the result in a new column as *new_last_name*.

NOTE: The three dots (...) indicate where to add the statement.

NOTE: SUBSTR takes in three arguments being column, starting_index, ending_index

```
SELECT
employee_id,
...
FROM
employee
ORDER BY
postal_code
```

Run Reset

What employee ID number is in row 8 of your query result?

NOTE: The query index starts at 1 not 0.

- 1
- 3
- 8
- 7

Correct
Feedback: The statement SUBSTR(last_name, 1, 3) AS new_last_name will retrieve the first 3 characters of each postal code and store the result in a new column as *new_last_name*. The complete query is SELECT employee_id, SUBSTR(last_name, 1, 3) AS new_last_name FROM employee ORDER BY postal_code. The SUBSTR function extracts a substring from a string. This function instructs the database to return 3 characters of each postal code, starting with the first character. The employee ID number 3 is in row 5 of your query result.

3.4 Module 4: Verify and report on your cleaning results

3.4.1 Manually cleaning data

Verifying the reporting results

This learning session emphasizes the importance of verifying and reporting in the data cleaning process. Verification ensures the accuracy of cleaned data by catching errors before analysis. Reporting maintains transparency, builds trust with stakeholders, and keeps everyone updated. It involves creating data-cleaning reports, documenting the process, and using changelogs. Changelogs track how a dataset evolves over time. The session highlights that these steps can prevent repeated mistakes and save time.

Cleaning and your data expectations

The video emphasizes the critical role of verification in data-cleaning efforts within analysis projects. Verification, likened to a "stamp of approval," ensures the accuracy and reliability of data, enabling reliable data-driven decision-making. The verification process begins with revisiting the original unclean dataset and comparing it with the cleaned version, using tools like conditional formatting and filters for efficient verification.

The video stresses the importance of maintaining a big-picture perspective during verification, ensuring alignment with the business problem and project goals. It acknowledges that projects can evolve over time and emphasizes the need for analysts to stay focused on the original objectives. A problem-first approach to analytics is highlighted, ensuring the data chosen for analysis directly addresses the business problem.

Three key considerations are outlined: focusing on the business problem addressed by the data, understanding the overarching project goal beyond mere data analysis, and assessing whether the collected and cleaned data can meet project objectives.

Analysts are advised to avoid becoming overly familiar with their data to prevent oversight or assumptions. Seeking fresh perspectives from teammates and obtaining feedback are recommended to enhance analysis accuracy.

The video underscores the importance of scrutinizing data for anomalies or suspicious patterns during the verification process. A hypothetical example involving an e-commerce company's customer satisfaction survey illustrates how data discrepancies may indicate underlying issues in the data-cleaning process.

The final step in data cleaning

The video continues the exploration of the verification process in data cleaning, emphasizing the need for thorough verification to ensure data is ready for analysis, akin to running tests on a car before it hits the road.

The initial verification step involves comparing the cleaned data with the original unclean dataset to identify common issues. Manual cleanup methods, such as removing extra spaces or unwanted characters, are discussed, and automated tools like TRIM and remove duplicates are introduced for automatic error resolution.

When errors persist and cannot be resolved manually or with automated tools, pivot tables are suggested as a data summarization tool. An example involving a party supply store's supplier data demonstrates how pivot tables can identify and correct errors like misspelled supplier names.

It also highlights two specific error resolution tools: the "Find and replace" tool, which replaces a specified term with another, and the pivot table, which counts occurrences and identifies data patterns. A step-by-step guide is provided for using these tools to correct a misspelled supplier name.

The video concludes with an introduction to the use of the CASE statement in SQL for error resolution. An example shows how a SQL query using the CASE statement corrects misspelled customer names, demonstrating the CASE statement's flexibility in handling multiple conditions.

Data-cleaning verification: A checklist

Correct the most common problems

Here is the list of your points converted to LaTeX itemize format:

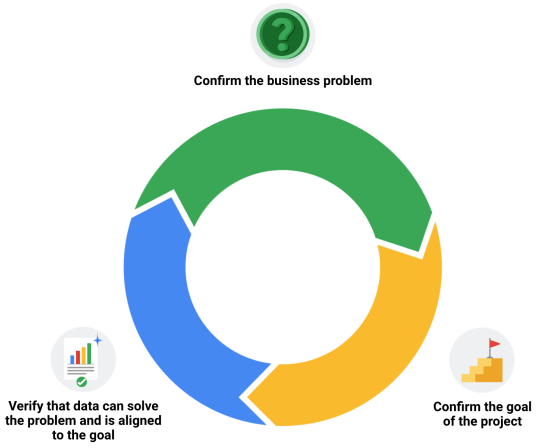
- **Sources of errors:** Did you use the right tools and functions to find the source of the errors in your dataset?
- **Null data:** Did you search for NULLs using conditional formatting and filters?
- **Misspelled words:** Did you locate all misspellings?
- **Mistyped numbers:** Did you double-check that your numeric data has been entered correctly?
- **Extra spaces and characters:** Did you remove any extra spaces or characters using the TRIM function?
- **Duplicates:** Did you remove duplicates in spreadsheets using the Remove Duplicates function or DISTINCT in SQL?
- **Mismatched data types:** Did you check that numeric, date, and string data are typecast correctly?
- **Messy (inconsistent) strings:** Did you make sure that all of your strings are consistent and meaningful?
- **Messy (inconsistent) date formats:** Did you format the dates consistently throughout your dataset?

- **Misleading variable labels (columns):** Did you name your columns meaningfully?
- **Truncated data:** Did you check for truncated or missing data that needs correction?
- **Business Logic:** Did you check that the data makes sense given your knowledge of the business?

Review the goal of your project

Once you have finished these data cleaning tasks, it is a good idea to review the goal of your project and confirm that your data is still aligned with that goal. This is a continuous process that you will do throughout your project– but here are three steps you can keep in mind while thinking about this:

- Confirm the business problem
- Confirm the goal of the project
- Verify that data can solve the problem and is aligned to the goal



3.4.2 Documenting results and the cleaning process

Capturing cleaning changes

This segment emphasizes the importance of documenting changes made during data cleaning. The video likens this to a crime TV show where forensic teams meticulously document evidence. Documentation serves three key purposes: it provides a reference for future error recovery, informs other users about changes made, and aids in evaluating data quality for analysis.

The video introduces the concept of a changelog, a tool used by data analysts to track modifications chronologically. This file, which can be implemented using spreadsheets or SQL, contains a list of modifications made to a project over time.

Practical guidance is provided on using Sheets' version history in spreadsheets to track changes, including the real-time tracking feature and the ability to view and revert to earlier versions. In SQL, the process of creating and viewing a changelog is discussed, emphasizing the importance of specifying changes made during query commits and utilizing query history.

Embrace changelogs

Keywords:

- ECOs (Engineering Change Orders): keep track of new product design details and proposed changes to existing products
- Changelogs: keep track of data transformation and cleaning

Automated version control takes you most of the way

Many software applications have built-in history tracking. For instance, Google Sheets allows you to check the version history of an entire sheet or an individual cell and revert to an earlier version if needed. To do this, you simply right-click the cell and select 'Show edit history'. Then, you can use the left or right arrow to navigate through the history.

Similarly, Microsoft Excel has a feature called 'Track Changes'. If this feature has been enabled for the spreadsheet, you can click on 'Review' and then under 'Track Changes', click the 'Accept/Reject Changes' option. This allows you to accept or reject any changes made.

In BigQuery, you can view the history to check what has changed. You can bring up a previous version without reverting to it and figure out what changed by comparing it to the current version. This feature is particularly useful in tracking the evolution of your data and ensuring accuracy.

Changelogs take you down the last mile

A changelog enhances your automated version history by providing a more comprehensive record of your work. It's where data analysts document all modifications made to the data. While version histories record what changes were made in a project, they don't explain why. Changelogs fill this gap by helping us understand the rationale behind changes. Changelogs don't have a fixed format and can be created in a blank document. However, if you're using a shared changelog, it's best to agree on a format with other data analysts.

Typically, a changelog records the following information:

- The data, file, formula, query, or other component that changed
- A description of what changed
- The date of the change
- The person who made the change

- The person who approved the change
- The version number
- The reason for the change

For instance, if you altered a formula in a spreadsheet to match another report and later discovered that the report had the wrong formula, an automated version history would help you undo the change. But if you also noted the reason for the change in a changelog, you could inform the creators of the report about the incorrect formula. If the change was made a long time ago, you might not remember who to contact. Fortunately, your changelog would have that information! By following up, you ensure data integrity beyond your project and demonstrate personal integrity as a trustworthy data handler. That's the power of a changelog!

Lastly, a changelog is crucial when numerous changes have been made to a spreadsheet or query. Imagine an analyst made four changes and wants to revert only the second change. Instead of using the undo feature three times (and losing the third and fourth changes), the analyst can undo just the second change and retain all the other changes. This example involves only four changes, but consider the importance of a changelog when there are hundreds of changes to track.

What also happens IRL (in real life)

In the context of a junior analyst making changes to an existing SQL query shared across a company, the company likely employs a version control system. This system impacts the modification process of a query as follows:

The company maintains official versions of significant queries in their version control system. The analyst ensures they are modifying the most recent version of the query, a process known as syncing. The analyst then makes a change to the query and may request a review of this change, informally or formally, such as asking a senior analyst to examine the change.

Once a reviewer approves the change, the analyst submits the updated version of the query to a repository in the company's version control system, a process known as a code commit. It is best practice to document precisely what the change was and why it was made. For instance, if a query that pulls daily revenue is updated to include revenue from a new product, Calypso, this would be documented.

After the change is submitted, everyone else in the company can access and use this new query when they sync to the most up-to-date queries stored in the version control system. If the query encounters a problem or business needs change, the analyst can undo the change using the version control system. The analyst can view a chronological list of all changes made to the query and who made each change. Then, after identifying their own change, the analyst can revert to the previous version.

The query reverts to its state before the analyst made the change, and everyone at the company sees this reverted, original query.

Why documentation is important

This segment emphasizes the importance of presenting the results of data cleaning efforts, drawing an analogy between data analysts and forensic scientists who present cleaned evidence in court. The importance of thorough documentation during the data cleaning process is highlighted, serving as a detailed record of all changes, additions, deletions, and errors.

The video introduces the concept of a changelog as a key form of chronological documentation, providing a real-time record of every adjustment made during the data cleaning process. This documentation is presented as a useful reference for future analysts dealing with similar datasets or errors.

The video revisits an example of an organization with duplicate membership instances to demonstrate how to document a data cleaning process. It suggests creating a document that outlines the steps taken and their effects, such as the removal of duplicate instances leading to a decrease in the total number of rows and a reduction in the total membership. For those using SQL, the video suggests including comments in the SQL statement as a more advanced documentation method.

The video emphasizes the importance of transparency in recording and sharing changelogs, with the aim of keeping all stakeholders informed. The objective is to show accountability for efficient data cleaning processes, thereby building trust as analysts who can accurately present evidence. The video ends by stating that resolving dirty data issues becomes a straightforward task when thorough documentation is maintained.

I am enrolling in a course called "Introduction to software engineering" at school. As a Product Manager for our group's project, I find this section very useful for my future career as a software engineer. I have learned about the importance of documentation in software development and feel familiar with the concept of writing documentation for the project. After this lecture, I will try to apply the knowledge I have learned to my group's project.

Feedback and cleaning

The video segment underscores the criticality of the data-cleaning process, with particular emphasis on verification, documentation, and reporting. These steps serve as concrete evidence to stakeholders, providing assurance that the data is both accurate and reliable, and that the cleaning process has been thoroughly executed and documented.

The video then shifts its focus to the importance of feedback on the presented evidence and its utilization for positive outcomes. It suggests that while clean data is essential, the data-cleaning process itself can unearth valuable insights that aid in business development. The video acknowledges common data errors, which could be due to human errors, flawed processes, or

system issues, as challenges. However, it highlights that consistent documentation and reporting of the cleaning process can illuminate the nature and severity of these error-generating processes.

Feedback obtained through reporting is emphasized as a transformative tool in the video. By identifying error patterns in data collection and entry procedures, organizations can leverage this feedback to prevent the recurrence of common errors. This could involve reprogramming data collection methods, altering survey questions, or even revising expectations and updating quality control procedures. In some instances, the video suggests that discussions with data engineers or data owners could be beneficial to ensure proper data integration and reduce the need for constant cleaning.

The video concludes by addressing the resolution of errors and inefficiencies in data collection, identified through the feedback process, which ultimately leads to trustworthy data for decision-making by stakeholders. It notes that reducing errors and enhancing efficiency in data collection can significantly boost the company's bottom line. The viewers are encouraged to acknowledge their newly acquired skills and look forward to further development in subsequent videos.

Advanced functions for speedy data cleaning

Keeping data clean and in sync with a source

The IMPORTRANGE function in Google Sheets and the Paste Link feature in Microsoft Excel provide efficient ways to insert data from one sheet into another, especially when dealing with large datasets. These methods offer advantages over manual copying and pasting by reducing the likelihood of errors and streamlining the process. This functionality is particularly valuable for data cleaning tasks, allowing users to selectively import relevant data for analysis while leaving out irrelevant information.

By utilizing these features, users can essentially "cancel noise" from their data, enabling a focused analysis on the most pertinent information to address specific problems. Moreover, this functionality proves beneficial for day-to-day data monitoring. Users can build tracking spreadsheets to share relevant data with others, and the synced nature ensures that the tracked data is automatically refreshed when the source data is updated.

In Google Sheets, the IMPORTRANGE function is employed for this purpose. This function allows users to specify a range of cells in another spreadsheet to duplicate within the current spreadsheet. However, it's crucial to note that access to the spreadsheet containing the data must be granted the first time the data is imported.

The example provided illustrates an analyst monitoring a fundraiser who needs to track and ensure the distribution of matching funds. They use IMPORTRANGE to pull matching transactions into a spreadsheet alongside individual donations. This allows them to track which donations eligible for matching funds still need processing. The dynamic nature of the data is accommodated by adjusting the range used by the function to import the most up-to-date information.

The provided URL is for syntax illustration only, and users are advised to replace it with the URL of their own spreadsheet, controlling access by clicking the "Allow access" button. The example emphasizes the iterative use of the IMPORTRANGE function, showcasing how the range can be adjusted to incorporate additional data as needed.

It's important to emphasize that the examples given are for illustrative purposes, and users should substitute their own URL and sheet name, along with the range of cells containing their specific data, when implementing these functions in their own spreadsheets.

Pulling data from other data sources

The QUERY function allows you to import and selectively filter data from another spreadsheet using SQL-like commands. This is particularly beneficial when dealing with large datasets, as it provides a quicker alternative to manual data filtering. For instance, you can utilize the QUERY function to generate a list of customers who made purchases in a specific month, all without modifying or duplicating your original data. This function also facilitates easy data retrieval for various months.

The syntax of the QUERY function is akin to that of the IMPORTRANGE function. You input the sheet name and the range of data you wish to query, then use SQL commands like SELECT to pinpoint the columns. Additional conditions can be appended after the SELECT statement using a WHERE clause. It's crucial to remember that all SQL code should be enclosed in quotation marks.

Both Google Sheets and Excel spreadsheets offer tools to connect to a data source and select tables. In both scenarios, you can ensure the imported data is verified and clean based on the query.

Analysts can leverage SQL to draw a specific dataset into a spreadsheet, and then employ the QUERY function to create different tabs or views of that dataset. Each tab can display a different subset of the data, such as sales data for a specific month or region. This exemplifies the effective synergy between SQL and spreadsheets.

I frequently utilize the IMPORTRANGE function to import data from different spreadsheets. However, the QUERY function is new to me. I've discovered that this function is incredibly beneficial and practical, particularly when extracting data from various data sources.

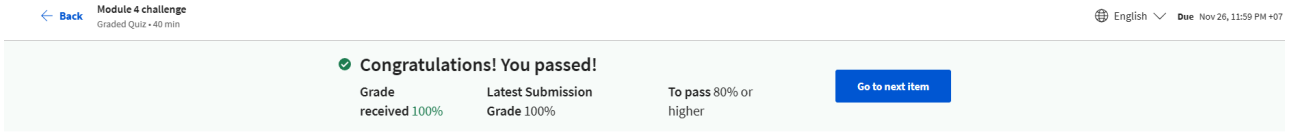
Filtering data to get what you want

The FILTER function operates entirely within a spreadsheet and doesn't necessitate the use of a query language. This function provides the capability to selectively view rows or columns in the source data based on specified conditions, allowing users to pre-filter data before conducting an analysis.

Compared to the QUERY function, the FILTER function may run faster. However, it's essential

to note that the QUERY function can be combined with other functions for more intricate calculations. For instance, the QUERY function can be integrated with functions like SUM and COUNT to perform data summarization tasks. In contrast, the FILTER function lacks this capacity for complex calculations and is primarily focused on the filtration of data based on user-defined conditions.

3.4.3 Module 4: Module challenge



1. Verification and reporting come directly before the data-cleaning process. 1 / 1 point

- ○ True
- ◉ False

Correct

2. What is the first step in the verification process? 1 / 1 point

- ○ Create a chronological list of modifications made to the data
- ◉ Compare cleaned data with the original, uncleaned dataset and compare it to what is there now
- ○ Determine the quality of the data
- ○ Inform others of your data-cleaning effort

Correct

3. While reviewing your coworker's data cleaning process, you find a few cases of trailing spaces in the data. What function can you use to remove these spaces? 1 / 1 point

- ◉ TRIM
- ○ CUT
- ○ DELETE
- ○ REMOVE TRAILING

Correct

4. A data analyst uses the COUNTA function to count which of the following? 1 / 1 point

- ○ The specific numbers in a dataset
- ○ The total number of headers in a specific range
- ○ The total number of entries in a changelog
- ◉ The total number of values within a specified range

Correct

5. Fill in the blank: A data analyst uses the CASE statement to consider one or more _____, then return a value. 1 / 1 point

- ○ changes
- ○ identifications
- ○ fields
- ◉ conditions

Correct

6. A data analyst uses a changelog to record how the data evolves while cleaning their data. What data cleaning best practice does this describe? 1 / 1 point

- ○ Illumination
- ○ Examination
- ◉ Documentation
- ○ Disclosure

Correct

7. At what point during the analysis process does a data analyst use a changelog? 1 / 1 point

- ◉ While cleaning the data
- ○ While visualizing the data
- ○ While gathering the data
- ○ While reporting the data

Correct

8. You use SQL to clean your data. You make comments whenever you modify your queries to keep track of any changes. What documentation will this practice help you create when you're done cleaning the data? 1 / 1 point

- ○ A new dataset
- ○ A database
- ○ A query repository
- ◉ A changelog

Correct

3.5 Module 5: Adding data to your resume

3.5.1 The data analyst hiring process

About the data-analyst hiring process

In our previous discussions, we delved into potential career paths upon completing your program and emphasized the importance of networking and establishing an online presence. Your active participation here reflects your commitment to shaping your future career. Now, our focus is on crafting a compelling resume. Whether you already have one or are considering a career switch, we'll explore the necessary adjustments. Before delving into resume building, we'll demystify the application process, ensuring you're well-prepared.

The journey involves understanding the nuances of creating a professional and tailored resume suitable for a data analyst role. We'll examine examples of effective resumes, providing practical insights. Following this, a self-analysis awaits, where we'll evaluate various types of data analyst positions.

The data-analyst hiring process

In this video, we're taking a step back from data analytics to explore what comes next after completing your program. Navigating the job search process can be challenging, but with the skills you're building as a data analyst, you're well-prepared for the journey ahead. Drawing from personal experience, reaching out to professionals to understand their career paths, companies, and roles can provide valuable insights. This parallels our current goal of offering you a glimpse into what to expect during your job search.

It's essential to recognize the uniqueness of each job search, influenced by factors such as location, field interests, and personal preferences in work environments. Emphasizing the individuality of this journey, we highlight common starting points, such as job-specific websites and company portals, where you can explore openings and set up alerts for relevant positions. Conducting thorough research on preferred companies and positions will guide you in tailoring your resume to align with specific requirements.

Professional networking, especially on platforms like LinkedIn, can play a pivotal role. Utilizing existing connections or reaching out to employees in your target companies may yield valuable insights or even lead to referrals. Acknowledging the inevitability of hearing "no" during the job search, we emphasize the importance of resilience and self-belief.

As your job search progresses, the initial point of contact may be a recruiter, who could be the bridge between you and potential employers. Maintaining professionalism during interactions is crucial, and leveraging technical terms demonstrates your competence. The hiring manager becomes a key player, evaluating your ability to contribute to the team. Researching hiring managers and asking insightful questions can enhance your candidacy.

Subsequent interviews allow future stakeholders and teammates to assess your suitability for the position. If all goes well, an official offer may follow, prompting the need for careful consideration and potential negotiation. Balancing your expectations with company values is essential in securing a competitive offer.

Upon accepting an offer, it's customary to provide at least a two-week notice to your current employer if applicable. This period allows for a smooth transition and a well-deserved break before embarking on your new role as a data analyst. This comprehensive overview aims to prepare you for the various stages of your job search, setting the stage for our upcoming discussion on resume building.

Creating a resume

Certainly, here's the information reorganized into paragraphs for a learning report:

—

In this section, we delve into the process of constructing an effective resume, treating it as a snapshot that encapsulates one's academic and professional journey. The objective is to provide a quick yet comprehensive overview of one's capabilities for hiring managers and recruiters. Emphasis is placed on brevity, recommending a one-page format with succinct bullet points, as busy professionals may only have time for a swift review.

The discussion extends to the utilization of templates, available on platforms such as Microsoft Word, Google Docs, and job search websites. These templates offer a structured framework with designated placeholders and design elements to enhance the visual appeal of the resume. The goal is to ensure professionalism, readability, and error-free content.

Contact information, typically placed at the top of the document, is a crucial starting point. Recommendations include using reliable and professional email addresses, preferably incorporating one's first and last name. The choice of formatting—whether emphasizing skills and qualifications over work history—depends on individual circumstances, such as career gaps or transitions.

The inclusion of a summary is optional but can be beneficial, particularly for those with non-traditional experience or undergoing career shifts. A well-crafted summary, limited to one or two sentences, highlights strengths and contributions to the prospective employer. The option of leaving a summary as a placeholder and refining it after completing other sections is also discussed, allowing for alignment with highlighted skills and experiences.

Work experience, including jobs, volunteer positions, and freelance work, is a pivotal section. The advice is to describe experiences in a way that aligns with the targeted position, emphasizing accomplishments and impact. Consideration is given to matching minimum and preferred qualifications listed in job descriptions, ensuring a competitive edge.

The importance of showcasing data analytics skills is emphasized, with a suggested formula

for descriptions: "Accomplished X as measured by Y, by doing Z." The relevance of the Google Data Analytics Professional Certificate and the application of newly gained skills are underscored, encouraging reflection on past experiences involving data analysis.

Education, including completion of courses such as the Google Data Analytics Professional Certificate, is recommended for inclusion. Furthermore, technical skills acquired, particularly in SQL, and language proficiencies can be highlighted in a dedicated section, augmenting one's qualifications.

The overall objective of the learning report is to guide individuals in creating a professional and compelling resume. As users progress, they will gain insights into further strategies to enhance their resumes and create a unique representation of their professional identity. The subsequent section of the learning report will delve into additional aspects that contribute to the distinctiveness of a resume.

3.5.2 Understand the element of a data analyst resume

Making your resume unique

In this segment, the focus is on refining a resume for data analytics jobs, recognizing the importance of showcasing clear communication skills. The introduction emphasizes the significance of effectively conveying analytical abilities, not just in performing analyses but also in articulating findings to diverse audiences. The primary audience is identified as hiring managers and recruiters, emphasizing the need for clarity and coherence in the resume.

The discussion begins with insights into the summary section, suggesting that it serves as an opportune space to highlight career transitions. The use of Problem-Action-Result (PAR) statements is introduced as a strategy for writing concise and clear descriptions. The PAR framework involves articulating the problem, the strategic action taken, and the achieved result. Examples illustrate how this approach can enhance the organization and effectiveness of job descriptions.

Moving to the skill section, the recommendation is to include relevant skills and qualifications acquired through the data analytics course. The emphasis is on demonstrating proficiency in key tools such as spreadsheets, SQL, Tableau, and the programming language R. The suggestion is made to create a dedicated section for programming languages, listing SQL and R, both integral components of the Google Data Analytics certificate. Additionally, highlighting proficiency in specific functions, packages, or formulas within these tools is encouraged.

A cautious note is provided, urging learners to accurately represent their skills and include these new skills only after completing the certificate. The promise is that applying the discussed ideas will set individuals apart from other candidates in the job market. The anticipation of completing the final course and having the opportunity to showcase skills through a case study, linkable on the resume, is presented as a valuable asset for impressing recruiters and hiring

managers.

The overall takeaway is that learners, upon completion of the program, will possess a well-crafted resume tailored for data analytics roles. The ability to continually update the resume as one pursues a career in data analytics is highlighted, promoting an ongoing commitment to professional development. The learning report will further elaborate on the strategies for incorporating experience into resumes, providing a comprehensive guide for learners in their job search endeavors.

CareerCon resources on YouTube

What is CareerCon?

CareerCon, an annual and complimentary online event hosted by Kaggle, is designed to assist budding data analysts in securing their initial role in the industry. The recorded sessions from CareerCon provide a wealth of direct insights and professional recommendations from leading data analysts and recruitment managers via lectures, coding training sessions, and resume guidance.

While the resources provided are primarily targeted at data scientists, the underlying principles and advice are also applicable to the career progression of data analysts.

CareerCon 2019 resourses

YouTube playlist, acknowledged that due to COVID-19 pandemic, CareerCon 2019 was the last event held and at the time of writing, there are no plans for future events.

Highlights from CareerCon 2018

- **How to build a compelling data science portfolio and resume:** A hiring manager from Quora reviews actual resumes from data science candidates and gives candid feedback on areas of improvement. Learn what to include and omit from your resume and portfolio as well as formatting tips. This offers a great firsthand look into what hiring managers are seeking when reviewing your resume and portfolio.
- **Overview of the Data Science Interview Process:** Hiring managers at Google discuss typical data science interviews, including the soft and hard skills you will want to prioritize. You will get a better sense of the interview process from both sides, and better prepare yourself for what to expect when interviewing for a data science role.
- **Live Breakdown of Common Data Science Interview Questions:** Watch a mock interview to see how a Kaggle data scientist answers questions during a data science interview. The video also includes live coding! This video is great preparation for some of the most commonly asked data science interview questions.
- **Am I a Good Fit? Identifying Your Best Data Science Job Opportunities:** Ever wonder

where you will fit in for your future career? This chat with Jessica Kirkpatrick, an intelligence manager, gives you a great breakdown of the different types of categories within the data science job market, the different types of job opportunities you may notice, and how you can frame previous work and skills from another career to fit into the data science job market.

- **Real Stories from a Panel of Successful Career Switchers:** Are you switching careers? Awesome! Learn from people who were in the same position as you and successfully switched their careers into data science. This panel discusses the different experiences in their careers and life that shifted them into the data science field.

3.5.3 Highlighting experience on resumes

Translating past work experience

1. **Communication Skills:**
   - Importance of articulating technical concepts to non-technical audiences.
   - Emphasis on past experiences in communication, such as presentations and interactions with various stakeholders.
   - Example: "Effectively implemented and communicated daily workflow, resulting in a 15% increase in productivity."
2. **Problem-Solving Skills:**
   - Data analysts are portrayed as problem-solvers, crucial in troubleshooting issues in databases or code.
   - Utilizing PAR (Problem, Action, Result) statements to showcase problem-solving abilities.
   - Example: Addressing a lack of daily workflow procedures, implementing them, and achieving a 15% productivity increase.
3. **Teamwork:**
   - Highlighting the importance of teamwork in data analysis, not only within the data team but for the entire company.
   - Stating how individual tasks contribute to the overall team and company success.
4. **Soft Skills:**
   - Soft skills, non-technical traits and behaviors, are crucial for success in data analysis.
   - Examples include being detail-oriented and demonstrating perseverance.
   - Relating soft skills to real-world examples from previous roles.

Conclusion

Emphasizes the power of transferable skills, especially soft skills, in enhancing the effectiveness of a data analyst resume. Encourages self-reflection on past roles to identify and highlight relevant transferable skills.

Adding professional skills to your resume

Common professional skills for entry-level data analysts

1. **Structured Query Language (SQL)**: SQL is a fundamental skill for entry-level data analyst roles. It facilitates interaction with databases, particularly for data retrieval. Each month, numerous data analyst job postings necessitate SQL knowledge, making it a common job function for data analysts.
2. **Spreadsheets**: Despite the prevalence of SQL, 62% of businesses continue to use spreadsheets for data insights. As a novice data analyst, your initial database might be in the form of a spreadsheet, a potent tool for data reporting and presentation. Hence, spreadsheet proficiency is crucial.
3. **Data Visualization Tools**: These tools simplify intricate data and make it visually comprehensible. Data analysts, after data collection and analysis, are responsible for presenting their findings in an easily understandable manner. Commonly used data analysis tools include Tableau, Microstrategy, Data Studio, Looker, Datarama, Microsoft Power BI, among others. Tableau, known for its user-friendliness, is essential for beginner data analysts. Furthermore, data analysis roles requiring Tableau are projected to increase by approximately 34.9% in the next decade.
4. **R or Python Programming**: Less than a third of entry-level data analyst roles require Python or R knowledge. While not necessary at the entry-level, proficiency in R or Python can be beneficial as you progress in your career.

Adding soft skills to your resume

Common soft skills for data analysts

- **Presentation skills**: the ability to structure and deliver data findings in a clear and simple way that suits the audience's needs and expectations.
- **Collaboration**: the ability to work effectively with different teams and stakeholders, both internal and external, and share ideas, insights, and feedback.
- **Communication**: the ability to obtain and convey data-related information in a language that is understandable and appropriate for the context.

- **Research**: the ability to conduct relevant and reliable research to analyze data and draw insights, and stay updated with industry trends and best practices.
- **Problem-solving skills**: the ability to identify and resolve errors and issues in data, databases, code, or data collection, and find alternative or creative solutions.
- **Adaptability**: the ability to adjust and cope with the ever-changing world of data, and work across multiple teams with different levels of needs and knowledge.
- **Attention to detail**: the ability to ensure accuracy and quality of data and code, and focus on the details that matter to the audience and the objective.

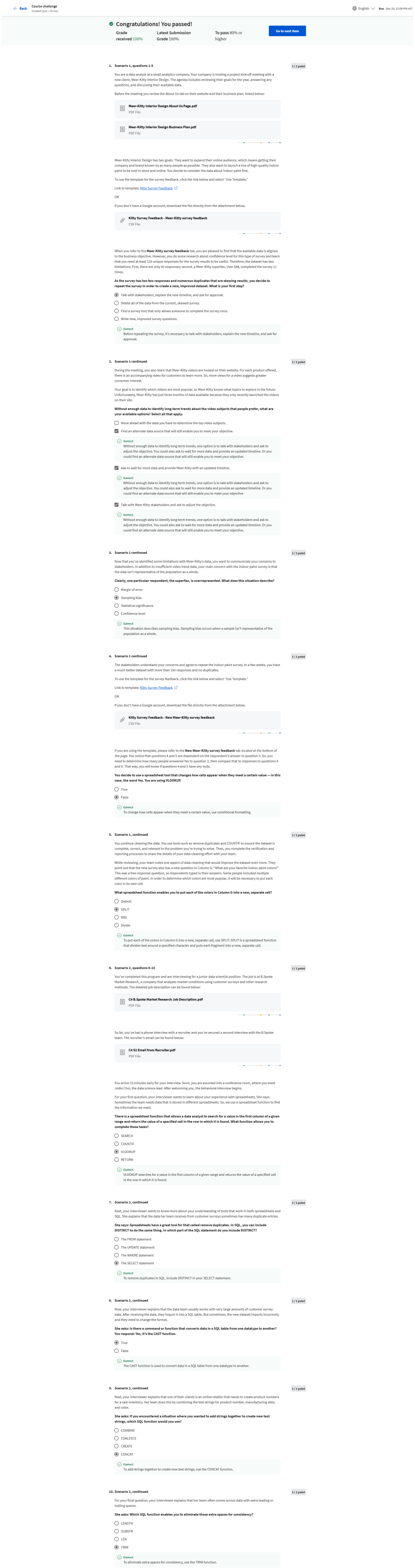
To showcase my soft skills on my resume, I learned three ways to do so:

1. Analyze my previous work experience and find opportunities to insert a soft skill. For example, if I worked in a restaurant, I could emphasize my communication and adaptability skills that I utilized to effectively function during peak hours.
2. Call attention to my problem-solving, presentation, research, and communication skills in previous projects or relevant coursework. For example, if I completed a data analysis project for a course, I could highlight how I researched, analyzed, and presented the data findings.
3. Add a mix of soft and professional skills in the skills or summary section of my resume. For example, I could list skills such as data analysis, SQL, Python, communication, collaboration, and problem-solving.

Conclusion

In this module, I gained valuable knowledge and skills on how to add soft skills to my resume as a data analyst. I learned about the importance and examples of soft skills, and how to demonstrate them on my resume. I believe that these skills will help me stand out from other candidates and impress potential employers.

3.6 Module 6: Course challenge



← Back Course challenge
Graded Quiz • 50 min

English Due Dec 10, 11:59 PM +07

Congratulations! You passed!

Grade received 100% | Latest Submission Grade 100% | To pass 80% or higher

Go to next item

1. Scenario 1, questions 1-5 — 1 / 1 point

You are a data analyst at a small analytics company. Your company is hosting a project kick-off meeting with a new client, Meer-Kitty Interior Design. The agenda includes reviewing their goals for the year, answering any questions, and discussing their available data.

Before the meeting you review the About Us tab on their website and their business plan, linked below:

Meer-Kitty Interior Design About Us Page.pdf
PDF File

Meer-Kitty Interior Design Business Plan.pdf
PDF File

Meer-Kitty Interior Design has two goals. They want to expand their online audience, which means getting their company and brand known by as many people as possible. They also want to launch a line of high-quality indoor paint to be sold in-store and online. You decide to consider the data about indoor paint first.

To use the template for the survey feedback, click the link below and select "Use Template."

Link to template: Kitty Survey Feedback

OR

If you don't have a Google account, download the file directly from the attachment below.

Kitty Survey Feedback - Meer-Kitty survey feedback
CSV File

When you refer to the **Meer-Kitty survey feedback** tab, you are pleased to find that the available data is aligned to the business objective. However, you do some research about confidence level for this type of survey and learn that you need at least 120 unique responses for the survey results to be useful. Therefore, the dataset has two limitations: First, there are only 40 responses; second, a Meer-Kitty superfan, User 588, completed the survey 11 times.

As the survey has too few responses and numerous duplicates that are skewing results, you decide to repeat the survey in order to create a new, improved dataset. What is your first step?

- (●) Talk with stakeholders, explain the new timeline, and ask for approval.
- ( ) Delete all of the data from the current, skewed survey.
- ( ) Find a survey tool that only allows someone to complete the survey once.
- ( ) Write new, improved survey questions.

Correct
Before repeating the survey, it's necessary to talk with stakeholders, explain the new timeline, and ask for approval.

2. Scenario 1 continued — 1 / 1 point

During the meeting, you also learn that Meer-Kitty videos are hosted on their website. For each product offered, there is an accompanying video for customers to learn more. So, more views for a video suggests greater consumer interest.

Your goal is to identify which videos are most popular, so Meer-Kitty knows what topics to explore in the future. Unfortunately, Meer-Kitty has just three months of data available because they only recently launched the videos on their site.

Without enough data to identify long-term trends about the video subjects that people prefer, what are your available options? Select all that apply.

- [ ] Move ahead with the data you have to determine the top video subjects.
- [x] Find an alternate data source that will still enable you to meet your objective.

Correct
Without enough data to identify long-term trends, one option is to talk with stakeholders and ask to adjust the objective. You could also ask to wait for more data and provide an updated timeline. Or you could find an alternate data source that will still enable you to meet your objective.

- [x] Ask to wait for more data and provide Meer-Kitty with an updated timeline.

Correct
Without enough data to identify long-term trends, one option is to talk with stakeholders and ask to adjust the objective. You could also ask to wait for more data and provide an updated timeline. Or you could find an alternate data source that will still enable you to meet your objective.

- [x] Talk with Meer-Kitty stakeholders and ask to adjust the objective.

Correct
Without enough data to identify long-term trends, one option is to talk with stakeholders and ask to adjust the objective. You could also ask to wait for more data and provide an updated timeline. Or you could find an alternate data source that will still enable you to meet your objective.

3. Scenario 1 continued — 1 / 1 point

Now that you've identified some limitations with Meer-Kitty's data, you want to communicate your concerns to stakeholders. In addition to insufficient video trend data, your main concern with the indoor paint survey is that the data isn't representative of the population as a whole.

Clearly, one particular respondent, the superfan, is overrepresented. What does this situation describe?

- ( ) Margin of error
- (●) Sampling bias
- ( ) Statistical significance
- ( ) Confidence level

Correct
This situation describes sampling bias. Sampling bias occurs when a sample isn't representative of the population as a whole.

4. Scenario 1 continued — 1 / 1 point

The stakeholders understand your concerns and agree to repeat the indoor paint survey. In a few weeks, you have a much better dataset with more than 150 responses and no duplicates.

To use the template for the survey feedback, click the link below and select "Use Template."

Link to template: Kitty Survey Feedback

OR

If you don't have a Google account, download the file directly from the attachment below.

Kitty Survey Feedback - New Meer-Kitty survey feedback
CSV File

If you are using the template, please refer to the **New Meer-Kitty survey feedback** tab *located at the bottom of the page*. You notice that questions 4 and 5 are dependent on the respondent's answer to question 3. So, you need to determine how many people answered Yes to question 3, then compare that to responses to questions 4 and 5. That way, you will know if questions 4 and 5 have any nulls.

You decide to use a spreadsheet tool that changes how cells appear when they meet a certain value — in this case, the word Yes. You are using VLOOKUP.

- ( ) True
- (●) False

Correct
To change how cells appear when they meet a certain value, use conditional formatting.

5. Scenario 1, continued — 1 / 1 point

You continue cleaning the data. You use tools such as remove duplicates and COUNTIF to ensure the dataset is complete, correct, and relevant to the problem you're trying to solve. Then, you complete the verification and reporting processes to share the details of your data-cleaning effort with your team.

While reviewing, your team notes one aspect of data cleaning that would improve the dataset even more. They point out that the new survey also has a new question in Column G: "What are your favorite indoor paint colors?" This was a free-response question, so respondents typed in their answers. Some people included multiple different colors of paint. In order to determine which colors are most popular, it will be necessary to put each color in its own cell.

What spreadsheet function enables you to put each of the colors in Column G into a new, separate cell?

- ( ) Delimit
- (●) SPLIT
- ( ) MID
- ( ) Divide

Correct
To put each of the colors in Column G into a new, separate cell, use SPLIT. SPLIT is a spreadsheet function that divides text around a specified character and puts each fragment into a new, separate cell.

6. Scenario 2, questions 6-10 — 1 / 1 point

You've completed this program and are interviewing for a junior data scientist position. The job is at B.Spoke Market Research, a company that analyzes market conditions using customer surveys and other research methods. The detailed job description can be found below:

C4 B.Spoke Market Research Job Description.pdf
PDF File

So far, you've had a phone interview with a recruiter and you've secured a second interview with the B.Spoke team. The recruiter's email can be found below:

C4 S2 Email from Recruiter.pdf
PDF File

You arrive 15 minutes early for your interview. Soon, you are escorted into a conference room, where you meet Jodie Choi, the data science lead. After welcoming you, the behavioral interview begins.

For your first question, your interviewer wants to learn about your experience with spreadsheets. She says: Sometimes the team needs data that is stored in different spreadsheets. So, we use a spreadsheet function to find the information we need.

There is a spreadsheet function that allows a data analyst to search for a value in the first column of a given range and return the value of a specified cell in the row in which it is found. What function allows you to complete these tasks?

- ( ) SEARCH
- ( ) COUNTIF
- (●) VLOOKUP
- ( ) RETURN

Correct
VLOOKUP searches for a value in the first column of a given range and returns the value of a specified cell in the row in which it is found.

7. Scenario 2, continued — 1 / 1 point

Next, your interviewer wants to know more about your understanding of tools that work in both spreadsheets and SQL. She explains that the data her team receives from customer surveys sometimes has many duplicate entries.

She says: Spreadsheets have a great tool for that called remove duplicates. In SQL, you can include DISTINCT to do the same thing. In which part of the SQL statement do you include DISTINCT?

- ( ) The FROM statement
- ( ) The UPDATE statement
- ( ) The WHERE statement
- (●) The SELECT statement

Correct
To remove duplicates in SQL, include DISTINCT in your SELECT statement.

8. Scenario 2, continued — 1 / 1 point

Now, your interviewer explains that the data team usually works with very large amounts of customer survey data. After receiving the data, they import it into a SQL table. But sometimes, the new dataset imports incorrectly and they need to change the format.

She asks: Is there a command or function that converts data in a SQL table from one datatype to another? You respond: Yes, it's the CAST function.

- (●) True
- ( ) False

Correct
The CAST function is used to convert data in a SQL table from one datatype to another.

9. Scenario 2, continued — 1 / 1 point

Next, your interviewer explains that one of their clients is an online retailer that needs to create product numbers for a vast inventory. Her team does this by combining the text strings for product number, manufacturing date, and color.

She asks: If you encountered a situation where you wanted to add strings together to create new text strings, which SQL function would you use?

- ( ) COMBINE
- ( ) COALESCE
- ( ) CREATE
- (●) CONCAT

Correct
To add strings together to create new text strings, use the CONCAT function.

10. Scenario 2, continued — 1 / 1 point

For your final question, your interviewer explains that her team often comes across data with extra leading or trailing spaces.

She asks: Which SQL function enables you to eliminate those extra spaces for consistency?

- ( ) LENGTH
- ( ) SUBSTR
- ( ) LEN
- (●) TRIM

Correct
To eliminate extra spaces for consistency, use the TRIM function.

4 Conclusion

Course 4, "Process Data from Dirty to Clean," has been a key stage in my development as a data analyst, focusing on the vital aspect of maintaining data quality and working with clean data. The course covered a wide range of topics that have greatly improved my skill set.

The methods to check data for integrity gave me a profound understanding of the importance of data quality in the analysis process. I learned how to deal with challenges from insufficient data, comprehend sample size, and avoid biases, which gave me the tools to handle real-world situations effectively.

The difference between clean and dirty data and the practice of data cleaning techniques in spreadsheets and other tools was a practical experience that enhanced my ability to transform raw data into a usable format. The use of SQL for data cleaning added a new dimension, allowing me to use structured query language to clean and manipulate data directly from databases.

The course not only stressed the technical aspects but also emphasized the importance of verifying and reporting cleaning results. This step ensures the reliability of the data used in the analysis, a fundamental aspect often neglected.

The optional module on adding data skills to my resume was a helpful addition, providing useful insights into the job application process. Creating a resume that highlights my strengths and relevant experience is a useful skill that will surely help me succeed in the field of data analytics.

The Course Challenge was the culmination of the course, a great opportunity to apply the key concepts learned throughout the modules. This practical exercise, involving real-world scenarios, will surely solidify my learning and prepare me for the challenges that await me in my data analytics journey.

As I reflect on the knowledge gained in Course 4, I am confident that I am better prepared to deal with the complexities of data processing, ensuring its cleanliness and quality. This course has not only added essential technical skills to my toolkit but has also emphasized the importance of attention to detail and the need for accuracy in every step of the data analysis process. I look forward to applying these skills in real-world situations and continuing my growth as a skilled data analyst.

In this learning report, I have intentionally left out certain activities, including Hands-on Activity, Ungraded plugins, and the Discussion forum. The reason for this omission is that the information provided by these activities was relatively basic for my current level of understanding, and I didn't deem it necessary to incorporate them into this report. Nonetheless, I have fully participated in and completed these activities.

Last words, I would like to give a special thanks to Mr.Đăng for encouraging me to write this

learning report. Thanks to you, I found the passion for learning MOOCs. I have received Google Data Analytics Professional Certificate and I am currently learning the Google IT Automation with Python Professional Certificate. I hope that I will have the opportunity to write more learning reports in the future.