Week I

Why data integrity is important

A strong analysis depends on the integrity of the data

Data integrity ✧.*
The accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle.

There's a chance that data could be **replicated, transferred,** or **manipulated**

Data replication ✧.*
The process of storing data in multiple locations

Data transfer ✧.*
The process of copying data from a storage device to memory, or from one computer to another

Data manipulation ✧.*
The process of changing data to make it more organized and easier to read

Other threats to data integrity
• Human error
• Viruses
• Malware
• Hacking
• System failures

More about data integrity and compliance

Data Constraint	Definition	Examples
Data type	Values must be of a certain type: date, number, percentage, Boolean, etc.	If the data type is a date, a single number like 30 would fail the constraint and be invalid
Data range	Values must fall between predefined maximum and minimum values	If the data range is 10-20, a value of 30 would fail the constraint and be invalid

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

Balancing objectives with data integrity

It's important to check that the data you use aligns with the business objective

Data duplicate ✧.*
Same customer's data showing up in more than one row.

Dealing with insufficient data

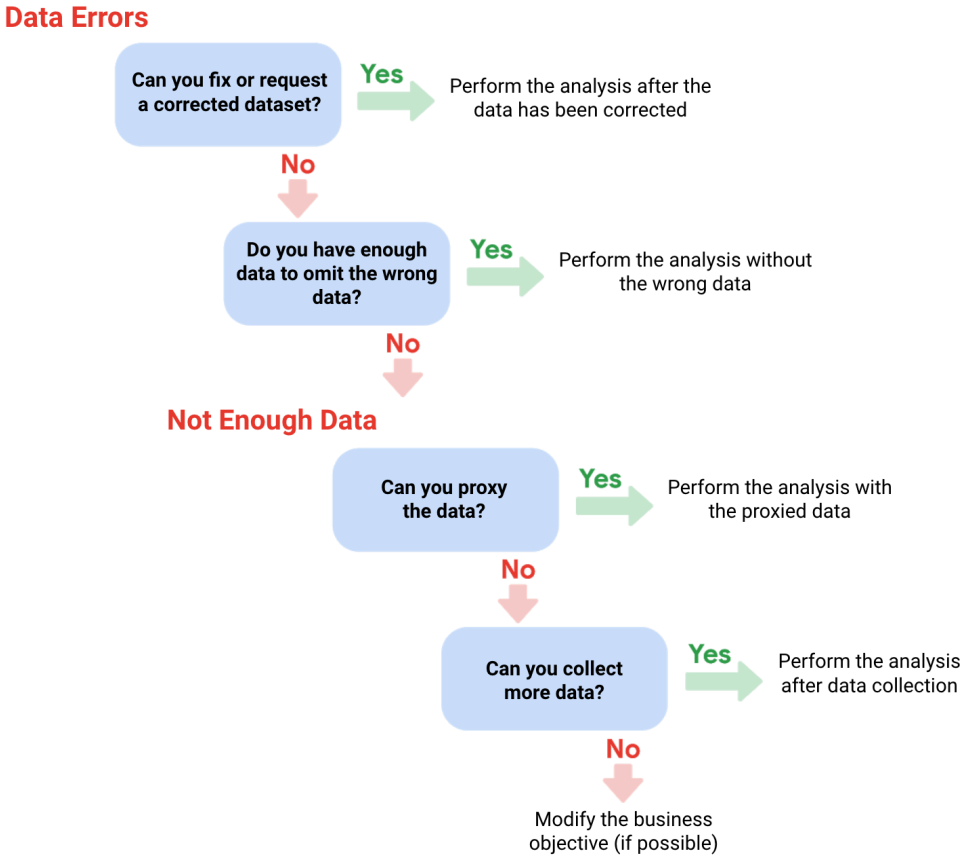
Types of insufficient data ✧.*

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

Ways to address insufficient data ✧.*

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

What to do when you find an issue with you data



Importance of sample size

Sample size ✧.*
a part of a population that is representative of the population.

Sampling bias ✧.*
when a sample isn't representative of the population as a whole

Random sampling ✧.*
A way of selecting a sample from a population so that every possible type of the sample has an equal chance of being chosen

Calculating sample size

Term	Definition	Example
Population	The entire group that you are interested in for your study.	All the employees in your company
Sample	A subset of your population.	A representative sample of your population
Margin of error	The difference between the sample's results and what the results would have been if you had surveyed the entire population.	If your sample's results are 50% and the margin of error is 5%, then the population's results could be anywhere from 45% to 55%.
Confidence level	How confident you are in the survey results.	A 95% confidence level means that you are 95% confident that the population's results are within the margin of error of the sample's results.
Confidence interval	The range of possible values that the population's result would be at the confidence level of the study.	If the sample's results are 50% and the margin of error is 5%, then the confidence interval is 45% to 55%.
Statistical significance	Whether your result could be due to random chance or not.	The greater the significance, the less likely it is that the result was due to chance.

Things to remember when determining the size of your sample

When figuring out a sample size, here are things to keep in mind:

- Don't use a sample size less than 30. It has been statistically proven that 30 is the smallest sample size where an average result of a sample starts to represent the average result of a population.
- The confidence level most commonly used is 95%, but 90% can work in some cases.

Increase the sample size to meet specific needs of your project:

- For a **higher** confidence level, use a larger sample size
- To **decrease** the margin of error, use a larger sample size
- For **greater** statistical significance, use a larger sample size

Note: Sample size calculators use statistical formulas to determine a sample size. More about these are coming up in the course! Stay tuned.

Why a minimum sample of 30?

This recommendation is based on the **Central Limit Theorem (CLT)** in the field of probability and statistics. As sample size increases, the results more closely resemble the normal (bell-shaped) distribution from a large number of samples. A sample of 30 is the smallest sample size for which the CLT is still valid. Researchers who rely on **regression analysis** – statistical methods to determine the relationships between controlled and dependent variables – also prefer a minimum sample of 30.

Still curious? Without getting too much into the math, check out these articles:

- Central Limit Theorem (CLT): This article by Investopedia explains the Central Limit Theorem and briefly describes how it can apply to an analysis of a stock index.
- Sample Size Formula: This article by Statistics Solutions provides a little more detail about why some researchers use 30 as a minimum sample size.

Sample sizes vary by business problem

Sample size will vary based on the type of business problem you are trying to solve.

For example, if you live in a city with a population of 200,000 and get 180,000 people to respond to a survey, that is a large sample size. But without actually doing that, what would an acceptable, smaller sample size look like?

Would 200 be alright if the people surveyed represented every district in the city?

Answer: It depends on the stakes.

- A sample size of 200 might be large enough if your business problem is to find out how residents felt about the new library
- A sample size of 200 might not be large enough if your business problem is to determine how residents would vote to fund the library

You could probably accept a larger margin of error surveying how residents feel about the new library versus surveying residents about how they would vote to fund it. For that reason, you would most likely use a larger sample size for the voter survey.

Larger sample sizes have a higher cost

You also have to weigh the cost against the benefits of more accurate results with a larger sample size. Someone who is trying to understand consumer preferences for a new line of products wouldn't need as large a sample size as someone who is trying to understand the effects of a new drug. For drug safety, the benefits outweigh the cost of using a larger sample size. But for consumer preferences, a smaller sample size at a lower cost could provide good enough results.

Knowing the basics is helpful

Knowing the basics will help you make the right choices when it comes to sample size. You can always raise concerns if you come across a sample size that is too small. A sample size calculator is also a great tool for this. Sample size calculators let you enter a desired confidence level and margin of error for a given population size. They then calculate the sample size needed to statistically achieve those results.
Refer to the Determine the Best Sample Size video for a demonstration of a sample size calculator, or refer to the Sample Size Calculator reading for additional information.

Using statistical power

Statistical power ✧.*
The probability of getting meaningful results from a test. You should know that statistical power is usually shown as a value out of one. So if your statistical power is 0.6, that's the same thing as saying 60%. In the milk shake ad test, if you found a statistical power of 60%, that means there's a 60% chance of you getting a statistically significant result on the ad's effectiveness.

Hypothesis testing ✧.*
A way to see if a survey or experiment has meaningful results

Statistically significant ✧.*
If a test is statistically significant, it means the results of the test are real and not an error caused by random chance. A 0.8 or 80% statistical power is typically considered the minimum for statistical significance.

Usually, you need a statistical power of at least zero point 8 or 80% to consider your results statistically significant

Consider what might prevent you from from getting statistically significant results.

- Are there restaurants running any other promotions that might bring in new customers?
- Do some restaurants have customers that always buy the newest item, no matter what it is?

- Do some locations have construction that recently started that would prevent customers from even going to the restaurant?

"Statistical power can be calculated and reported for a completed experiment to comment on the confidence one might have in the conclusions drawn from the results of the study. It can also be used as a tool to estimate the number of observations or sample size required in order to detect an effect in an experiment."

What to do when there is no data

Proxy data examples

Sometimes the data to support a business objective isn't readily available. This is when proxy data is useful. Take a look at the following scenarios and where proxy data comes in for each example:

Business scenario	How proxy data can be used
A new car model was just launched a few days ago and the auto dealership can't wait until the end of the month for sales data to come in. They want sales projections now.	The analyst proxies the number of clicks to the car specifications on the dealership's website as an estimate of potential sales at the dealership.
A brand new plant-based meat product was only recently stocked in grocery stores and the supplier needs to estimate the demand over the next four years.	The analyst proxies the sales data for a turkey substitute made out of tofu that has been on the market for several years.
The Chamber of Commerce wants to know how a tourism campaign is going to impact travel to their city, but the results from the campaign aren't publicly available yet.	The analyst proxies the historical data for airline bookings to the city one to three months after a similar campaign was run six months earlier.

Open (public) datasets

If you are part of a large organization, you might have access to lots of sources of data. But if you are looking for something specific or a little outside your line of business, you can also make use of open or public datasets. (You can refer to this Medium article for a brief explanation of the difference between open and public data.)

Here's an example. A nasal version of a vaccine was recently made available. A clinic wants to know what to expect for contraindications, but just started collecting first-party data from its patients. A **contraindication** is a condition that may cause a patient not to take a vaccine due to the harm it would cause them if taken. To estimate the number of possible contraindications, a

data analyst proxies an open dataset from a trial of the injection version of the vaccine. The analyst selects a subset of the data with patient profiles most closely matching the makeup of the patients at the clinic.

There are plenty of ways to share and collaborate on data within a community. Kaggle (kaggle.com) which we previously introduced, has datasets in a variety of formats including the most basic type, Comma Separated Values (CSV) files.

SV, JSON, SQLite, and BigQuery datasets

- CSV: Check out this Credit card customers dataset, which has information from 10,000 customers including age, salary, marital status, credit card limit, credit card category, etc. (CC0: Public Domain, Sakshi Goyal).
- JSON: Check out this JSON dataset for trending YouTube videos (CC0: Public Domain, Mitchell J).
- SQLite: Check out this SQLite dataset for 24 years worth of U.S. wildfire data (CC0: Public Domain, Rachael Tatman).
- BigQuery: Check out this Google Analytics 360 sample dataset from the Google Merchandise Store (CC0 Public Domain, Google BigQuery).

Refer to the Kaggle documentation for datasets for more information and search for and explore datasets on your own at kaggle.com/datasets.
As with all other kinds of datasets, be on the lookout for duplicate data and 'Null' in open datasets. Null most often means that a data field was unassigned (left empty), but sometimes Null can be interpreted as the value, 0. It is important to understand how Null was used before you start analyzing a dataset with Null data.

Determine the best sample size

Confidence level ✧.*
The probability that your sample size accurately reflects the greater population.

Having a 99% confidence level is ideal, but most industries hope for at least a 90% or 95% percent confidence level

Margin error ✧.*
Tells you how close your sample size results are to what your results would be if you use the entire population that your sample size represents.

Sample size calculator

A **sample size calculator** tells you how many people you need to interview (or things you need to test) to get results that represent the target population. Let's review some terms you will come across when using a sample size calculator:

- **Confidence level**: The probability that your sample size accurately reflects the greater population.
- **Margin of error**: The maximum amount that the sample results are expected to differ from those of the actual population.
- **Population**: This is the total number you hope to pull your sample from.
- **Sample**: A part of a population that is representative of the population.
- **Estimated response rate**: If you are running a survey of individuals, this is the percentage of people you expect will complete your survey out of those who received the survey.

What to do with the results

After you have plugged your information into one of these calculators, it will give you a recommended sample size. Keep in mind, the calculated sample size is the **minimum** number to achieve what you input for confidence level and margin of error. If you are working with a survey, you will also need to think about the estimated response rate to figure out how many surveys you will need to send out. For example, if you need a sample size of 100 individuals and your estimated response rate is 10%, you will need to send your survey to 1,000 individuals to get the 100 responses you need for your analysis.

Evaluate the reliability of your data

Margin error ✧.*

The maximum amount that the sample results are expected to differ from those of the actual population. The closer to zero the margin of error, the closer your results from your sample would match results from the overall population.

To calculate margin of error you need:

- Population size
- Sample size
- Confidence level

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.

Week II

Clean it up!

#1 cause of poor quality data = human error

Dirty data ✧.*

Data that is incomplete, incorrect, or irrelevant to the problem you're trying to solve

Clean data ✧.*

Data that is complete, correct, and relevant to the problem you're trying to solve

Why data cleaning is important

Data engineers ✧.*

Transform data into a useful format for analysis and give it a reliable infrastructure. This means they develop, maintain, and test databases, data processors and related systems.

Data warehousing specialists ✧.*

develop processes and procedures to effectively store and organize data. They make sure that data is available, secure, and backed up to prevent loss.

Null ✧.*

An indication that a value does not exist in a dataset

What is dirty data?

Duplicate data

Outdated data

Incomplete data

Incorrect/inaccurate data

Inconsistent data

Type of dirty data	Description	Possible causes	Potential harm to businesses
Duplicate data	Any data record that shows up more than once	Manual data entry, batch data imports, or data migration	Skewed metrics or analyses, inflated or inaccurate counts or predictions, or confusion during data retrieval
Outdated data	Any data that is old which should be replaced with newer and more accurate information	People changing roles or companies, or software and systems becoming obsolete	Inaccurate insights, decision-making, and analytics
Incomplete data	Any data that is missing important fields	Improper data collection or incorrect data entry	Decreased productivity, inaccurate insights, or inability to complete essential services

Incorrect/inaccurate data	Any data that is complete but inaccurate	Human error inserted during data input, fake information, or mock data	Inaccurate insights or decision-making based on bad information resulting in revenue loss
Inconsistent data	Any data that uses different formats to represent the same thing	Data stored incorrectly or errors inserted during data transfer	Contradictory data points leading to confusion or inability to classify or segment customers

Recognize and remedy dirty data

Field length ✧.*
A tool for determining how many characters can be keyed into a field

Data validation ✧.*
A tool for checking the accuracy and quality of data before adding or importing it

Data Integrity/Clean and Dirty

Principles of data integrity

Accuracy ✧.*
The degree of conformity of a measure to a standard or a true value

Completeness ✧.*
The degree to which all required measures are known

Consistency ✧.*
The degree to which a set of measures is equivalent across systems

Validity ✧.*
The concept of using data integrity principles to ensure measures conform to defined business rules or constraints

Data-cleaning tools and techniques

- **Remove unwanted data**

Before removing unwanted data, it's always a good practice to make a copy of the data set.

Irrelevant data ✧.*
data that doesn't fit the specific problem that you're trying to solve, also needs to be removed.

- **Remove extra spaces and blanks**
- **Fixing misspellings**
- **Inconsistent capitalization**
- **Incorrect punctuation and other typos**
- **Remove formatting**

Cleaning data from multiple sources

Merger ✧.*
An agreement that unites two organizations into a single new one

Data merging ✧.*
The process of combining two or more datasets into a single dataset

Compatibility ✧.*
How well two or more datasets are able to work together

- Do I have all the data I need?
- Does the data I need exist within these datasets?
- Does the data need to be cleaned, or are they ready for me to use?
- Are the datasets cleaned to the same standard?

Common data-cleaning pitfalls

Data-cleaning features in spreadsheets

Conditional formatting ✧.*
A spreadsheet tool that changes how cells appear when values meet specific conditions

Remove duplicates ✧.*
A tool that automatically searches for and eliminates duplicate entries from a spreadsheet

Text string ✧.*
A group of characters within a cell, most often composed of letters

Length ✧.*
Number of characters in a text string.

String ✧.*
Smaller subset of a text string

Split ✧.*
A tool that divides text around a specified character and puts each fragment into a new, separate cell

Specified text separator ✧.*
Also known as delimiter, a character that indicates the beginning or end of a data item.

Concatenate ✧.*
A function that joins multiple text strings into a single string

Optimize the data-cleaning process

COUNTIF ✧.*
A function that returns the number of cells that match a specified value
SYNTAX: =COUNTIF(range, "value")

Syntax ✧.*
A predetermined structure that includes all required information and its proper placement

LEN ✧.*
A function that tells you the length of a text string by counting the number of characters it contains SYNTAX: =LEN(range)

LEFT ✧.*
A function that gives you a set number of characters from the left side of a text string

SYNTAX: =LEFT(range, number of characters)

RIGHT ✧.*
A function that gives you a set number of characters from the right side of a text string
SYNTAX:=RIGHT(range, number of characters)

MID ✧.*
A function that gives you a segment from the middle of a text string
SYNTAX:=MID(range, reference starting point, number of middle characters)

CONCATENATE ✧.*
SYNTAX:=CONCATENATE(item 1, item 2)

TRIM ✧.*
A function that removes leading, trailing, and repeated spaces in data
SYNTAX:=TRIM(range)

Different data perspectives

Different methods that data analysts use to look at data differently and how that leads to more efficient and effective data cleaning. Some of these methods include sorting and filtering, pivot tables, a function called VLOOKUP, and plotting to find outliers.

VLOOKUP ✧.*
Stands for vertical lookup. A function that searches for a certain value in a column to return a corresponding piece of information
SYNTAX:=VLOOKUP(data to look up, 'where to look'!Range, column, false)

Even more data-cleaning techniques

Data mapping ✧.*
The process of matching fields from one data source to another

Compatibility ✧.*
How well two or more datasets are able to work together

Schema ✧.*
A way of describing how something is organized

Primary key ✧.*
an identifier that references a column in which each value is unique.

Foreign key ✧.*
a field within a table that's a primary key in another table.

Week III

Understanding SQL capabilities

SQL ✧.*

a structured query language that analysts use to work with databases. Data analysts usually use SQL to deal with large datasets because it can handle huge amounts of data.

*In **1970**, Edgar F.Codd developed the theory about relational databases.*At the time IBM was using a relational database management system called System R. Well, IBM computer scientists were trying to figure out a way to manipulate and retrieve data from IBM System R. Their first query language was hard to use. So they quickly moved on to the next version, SQL. In **1979**, after extensive testing SQL, now just spelled S-Q-L, was released publicly. By **1986**, SQL had become the standard language for relational database communication, and it still is.

Using SQL as a junior data analyst

Features of Spreadsheets	Features of SQL Databases
Smaller data sets	Larger datasets
Enter data manually	Access tables across a database
Create graphs and visualizations in the same program	Prepare data for further analysis in another software
Built-in spell check and other useful functions	Fast and powerful functionality
Best when working solo on a project	Great for collaborative work and tracking queries run by all users

Widely used SQL queries

use **SELECT** to specify exactly what data we want to interact with in a table.

If we combine SELECT with **FROM**, we can pull data from any table in this database as long as they know what the columns and rows are named.

Spreadsheets versus SQL

Spreadsheets	SQL
Generated with a program	A language used to interact with database programs
Access to the data you input	Can pull information from different sources in the database
Stored locally	Stored across a database
Small datasets	Larger datasets
Working independently	Tracks changes across team
Built-in functionalities	Useful across multiple programs

SQL dialects and their uses

- LearnSQL's blog, **What Is a SQL Dialect, and Which One Should You Learn?**
- Software Testing Help's article, **Differences Between SQL Vs MySQL vs SQL Server**
- Datacamp's blog, **SQL Server, PostgreSQL, MySQL... what's the difference? Where do I start?** Note that there is an error in this blog article. The comparison table incorrectly states that SQlite uses subqueries instead of window functions. Refer to the **SQLite Window Functions** documentation for proper clarification.
- SQL Tutorial's tutorial, **What is SQL**

Widely used SQL queries

Inputting new information ✧.*

INSERT INTO - *where we want to insert*

(specifying which column) - typing in parenthesis

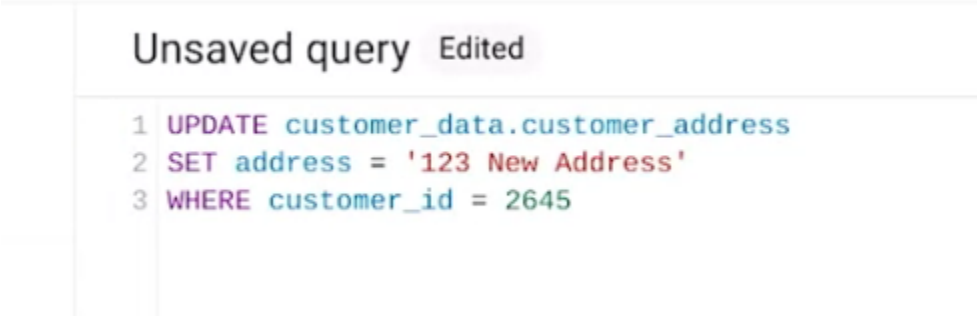
VALUES *- what values we want to insert*

Updating information ✧.*

UPDATE - what table we want to update

SET - what value we trying to change (new)

WHERE - where specifically we try to change the data. It *should be contained between two ticks, i.e. 'value*



DROP TABLE IF EXISTS ✧.*

if you're creating lots of tables within a database, you'll want to use the DROP TABLE IF EXISTS statement to clean up after yourself.

Cleaning string variables using SQL

Distinct ✧.*

Includes DISTINCT in your SELECT statement removes duplicates

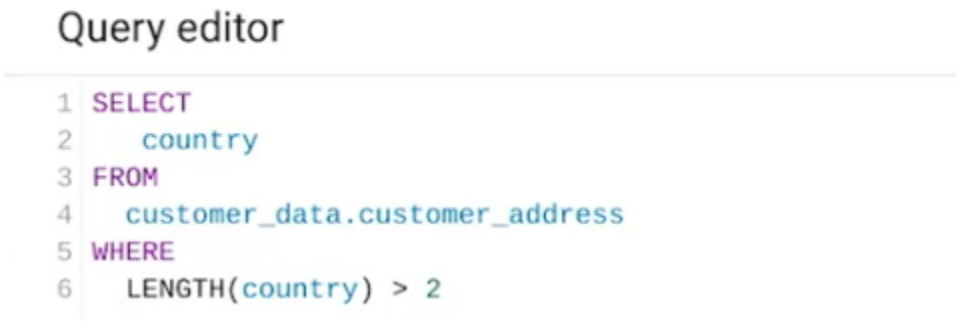
Len ✧.*

(Length) Use to double check our string variables as consistent.

LENGTH after the SELECT and place which column we want to check.

as is the label of the column we want to place the result of the length.

To find which of the data is not uniform ✧.*



Substring() ✧.*

To pull only two letters without changing the dataset.

SUBSTR(column error, start, end including start)

Trim ✧.*

Removes any spaces.

TRIM(column we want to change = ‘what data specifically‘)

Advanced data cleaning function, p1

CAST ✧.*

Can be used to convert anything from one data type to another

- **DESC** - descending
- **ASCE** - ascending

Float ✧.*

A number that contains a decimal

Typecasting ✧.*

Converting data from one type to another

Unsaved query Edited

```
1 SELECT
2   CAST(purchase_price AS FLOAT64)
3 FROM
4   customer_data.customer_purchase
5 ORDER BY
6   CAST(purchase_price AS FLOAT64) DESC
```

Changing the data type

Advanced data cleaning function, p2

To show only the date and not the time in the dateandtime format ✧.*

```
1 SELECT
2   CAST(date AS date) AS date_only,
3   purchase_price
4 FROM
5   customer_data.customer_purchase
6 WHERE
7   date BETWEEN '2020-12-01' AND '2020-12-31'
```

Query results SAVE RESULTS E

Query complete (0.7 sec elapsed, 436 B processed)

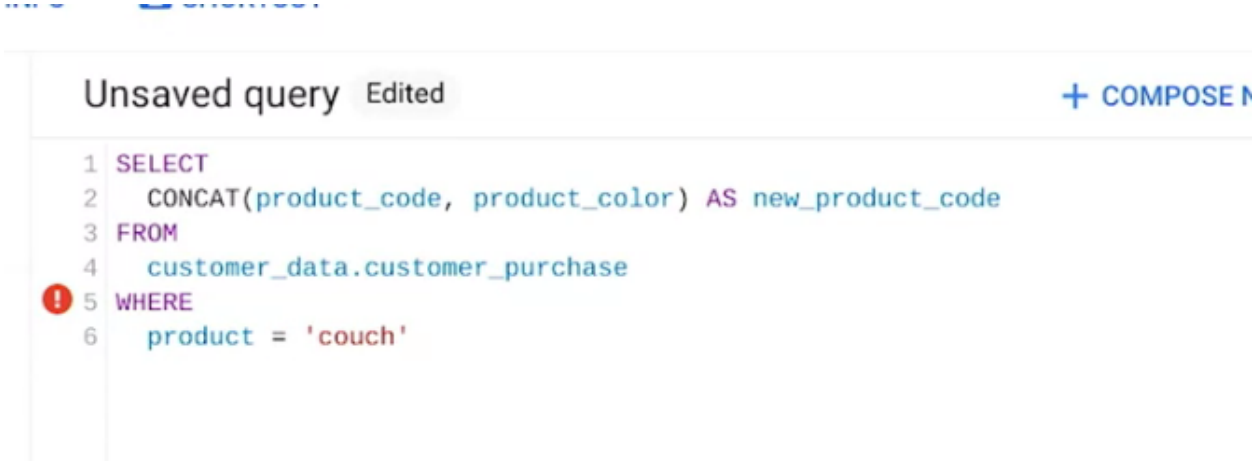
Job information | Results | JSON | Execution details

Row	date_only	purchase_price
1	2020-12-12	13.99
2	2020-12-28	27.98
3	2020-12-30	269.55
4	2020-12-28	229.95

3:08 / 8:55

CONCAT() ✧.*

Adds strings together to create new text strings that can be used as unique keys



Separate product by color

CONCAT(column, column) as new_product_column

COALESCE() ✧.*

Can be used to return non-null values in a list. COALESCE can save you time when you're making calculations too by skipping any null values and keeping your math correct.

COALESCE(first column to check, second) AS new name of field

Data cleaning with SQL

Data-cleaning SQL functions

Matching exercise

Definition	Query
Return a limited number of characters to create substrings from longer strings of text	SUBSTR()
Return the length of a string of text by counting the number of characters it contains	LENGTH()/LEN()
Pull data from a specific place in a table, typically a table column	SELECT FROM WHERE
Pull data from any table in a database	SELECT FROM

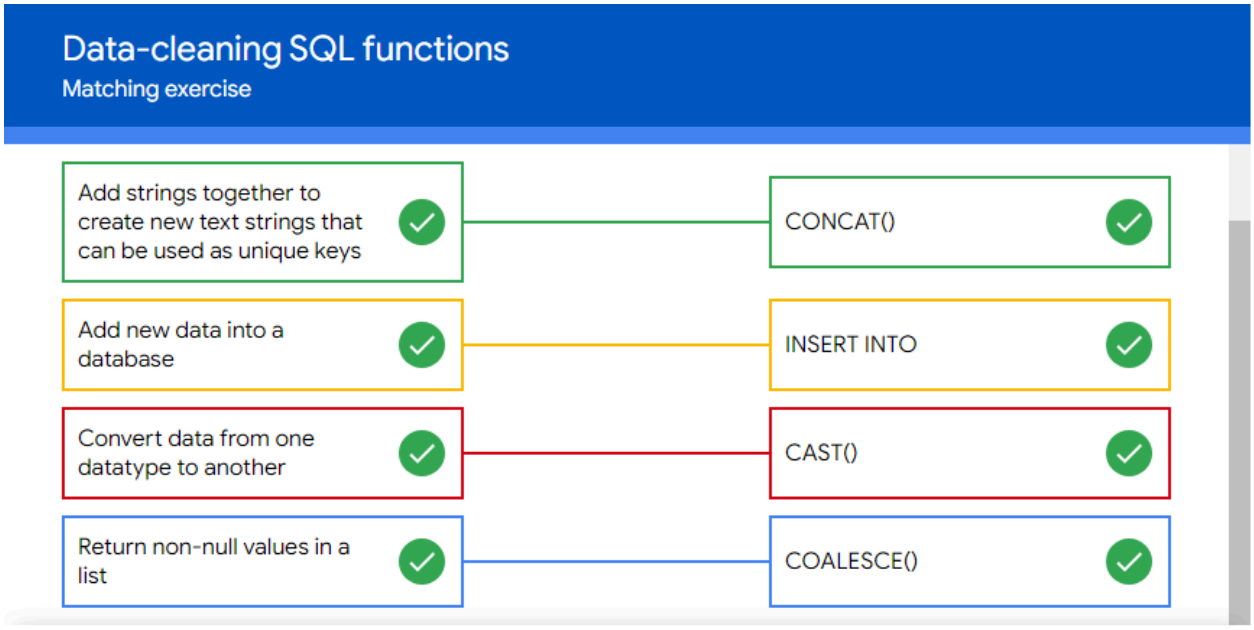
Data-cleaning SQL functions

Matching exercise

definition. Then, select a query to find out if it's a match.

Select **Next** to continue

Definition	Query
Change existing data in a database	UPDATE
Remove leading, trailing, and repeated spaces in data	TRIM()
Remove data from a database	DELETE



Week IV

Verifying and reporting result

Verification ✧.*

A process to confirm that a data-cleaning effort was well-executed and the resulting data is accurate and reliable

Changelog ✧.*

A file containing a chronologically ordered list of modifications made to a project

Cleaning and your data expectations

See the big picture when verifying data-cleaning ✧.*

1. **Consider the business problem**

Taking a problem-first approach to analytics is essential at all stages of any project. You need to be certain that your data will actually make it possible to solve your business problem.

2. **Consider the goal**

t's not enough just to know that your company wants to analyze customer feedback about a product. What you really need to know is that the goal of getting this feedback is to make

improvements to that product. On top of that, you also need to know whether the data you've collected and cleaned will actually help your company achieve that goal.

3. **Consider the data**

That means thinking about where the data came from and testing your data collection and cleaning processes.

The final step in data cleaning

Pivot table ✧.*

a data summarization tool that is used in data processing. Pivot tables sort, reorganize, group, count, total or average data stored in a database.

Find and replace ✧.*

A tool that looks for a specified search term in a spreadsheet and allows you to replace it with something else

COUNTA ✧.*

A function that counts the total number of values within a specified range

COUNT ✧.*

only counts the numerical values within a specified range.

CASE statement ✧.*

The CASE statement goes through one or more conditions and returns a value as soon as a condition is met (use in SQL)

Changing typos in SQL

Unsaved query Edited

```
1  SELECT
2    customer_id,
3    CASE
4      WHEN first_name = 'Tnoy' THEN 'Tony'
5      WHEN first_name = 'Tmo' THEN 'Tom'
6      WHEN first_name = 'Rachle' THEN
7      ELSE first_name
8      END AS cleaned_name
9  FROM
10   customer_data.customer_name
```

Data-cleaning verification: A checklist

Correct the most common problems

Make sure you identified the most common problems and corrected them, including:

- **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

Capturing cleaning changes

Documentation ✧.*

The process of tracking changes, additions, deletions, and errors involved in your data-cleaning effort

- Recover data-cleaning errors

- Inform other users of changes
- Determine the quality of data

Changelog ✧.*
A file containing a chronologically ordered list of modifications made to a project

Feedback and cleaning

Common data errors
• Human error in data entry
• Flawed processes
• System issues

Advanced functions for speedy data cleaning

Function	Syntax (Google Sheets)	Menu Options (Microsoft Excel)	Primary Use
IMPORTRANGE	=IMPORTRANGE(spreadsheet_url , range_string)	Paste Link (copy the data first)	Imports (pastes) data from one sheet to another and keeps it automatically updated.
QUERY	=QUERY(Sheet and Range, "Select *")	Data > From Other Sources > From Microsoft Query	Enables pseudo SQL (SQL-like) statements or a wizard to import the data.
FILTER	=FILTER(range, condition1, [condition2, ...])	Filter (conditions per column)	Displays only the data that meets the specified conditions.

Week V

The data analyst job-application process

1. Check out available jobs
    - Job sites
    - Company websites

Once you find a few that you like, do some research to learn more about the companies and the details about the specific positions you'll be applying for. Then you can update your resume or create a new one. You'll want it to be specific and reflect what each company is looking for. But you can definitely have a master resume that you tweak for each position.

2. Keep building and refreshing your job profile

- Be professional and personable.
- Using technical terms like "SQL" and "clean data" will show recruiters that you know what you're doing.
- Recruiters probably won't go into too much detail about the ins and outs. But they want to see that you know what you're talking about. They might also give you prep materials or other recommendations.

3. Hiring manager

The hiring manager's job is to evaluate whether you have the ability to do the work and whether you'd be a good fit for their team. Your job is to convince them that yes, you do, and yes, you would be.

A good thing you can do here is use LinkedIn or other professional sites to research the hiring managers or even other analysts who have a similar role to the one you're applying for. The more information you have about the job, the better your chances of actually getting it. You should also use this opportunity to ask lots of questions to help you figure out if the company's a good fit for you. You can do this when you talk to recruiters too.

4. Make sure it’s competitive offer

Remember, if they reach out to you with an offer, that means they want you as much as you want them. If you're interviewing at other places, you can leverage this to figure out if negotiating for a more competitive offer is possible. You should also research salaries, benefits, vacation time, and any other factors that are important to you for similar jobs.

Creating a resume

1. You want your resume to be a snapshot of all that you've done both in school and professionally.
2. **Be brief** - Try to keep everything in one page and each description to just a few bullet points. Two to four bullet points is enough but remember to keep your bullet points concise. Sticking to one page will help you stay focused on the details that best reflect who you are or who you want to be professionally. One page might also be all that hiring managers and recruiters have time to look at.

- most have contact information at the top of the document. This includes your name, address, phone number, and email address. If you have multiple email addresses or phone numbers, use the ones that are most reliable and sound professional. It's also great if you can use your first and last name in your email address
- A format that focuses more on skills and qualifications nd less on work history is great for people who have gaps in their work history It's also good for those who are just starting out their career or making a career change, and that might be you.
- If you do want to highlight your work history, feel free to include details of your work experience starting with your most recent job. If you've had lots of jobs that are related to a new position you're applying for, this format make sense.
- If you're editing a resume you already have, you can keep it in the same format and adjust the details.
- If you're starting a new one or building a resume for the first time, choose the format that makes the most sense for you.

- **Summary**. If you decide to include a summary, keep it to one or two sentences that highlight your strengths and how you can help the company you're applying to. You'll also want to make sure your summary includes positive words about yourself, like dedicated and proactive. You can support those words with data, like the number of years you've worked or the tools you're experienced in like SQL and spreadsheets.

A summary might start off with something like a hardworking customer service representative with over five years of experience.

"An entry-level data analytics professional recently completed the Google Data Analytics Professional Certificate."

Outside of jobs with other companies, you could also include volunteer positions you've had and any freelance or side work you've done. The key here is the way in which you describe these experiences. Try to describe the work you did in a way that relates to the position you're applying for.

- Checking out preferred qualifications, which lots of job descriptions also include. These aren't required, but every additional qualification you match makes you a more competitive candidate for the role. Including any part of your skills and experience that matches a job description will help your resume rise above the competition.

Making your resume unique

- Make sure you're a clear communicator - explain in a clear and direct way. Be coherent.
- Summary:
    - If transitioning: Transitioning from a career in the auto industry and seeking a full-time role in the field of data analytics

Use P.A.R - problem, action, resolve.

Instead: Was responsible for writing two blogs a month
Use: Earned little-known website over 2,000 organic clicks through strategic blogging

Problem: Earned little-known website
Strategic Action: strategic blogging
Resolution: 2,000 organic clicks

Speaking of the skill section, make sure you include any skills and qualifications you've acquired through this course and on your own. You don't need to be super technical. But talking about your experience with **spreadsheets, SQL, Tableau,** and **R.**

If you're listing qualifications or skills, you might include a spot for programming languages and then list SQL and R

Translating past work experience

Transferable skills ✧.*
Skills and qualities that can transfer from one job or industry to another

When job descriptions say they want strong communication skills for a data analyst, it usually means they want someone who can speak about what they do to people who aren't as technical or analytical.

In your work history section, you can highlight how your effective communication skills have helped you. You can also refer to specific presentations you've made and the outcomes of those presentations, and you can even include the audience for your presentations, especially if you present it to large groups or people in senior positions.

After listing job details, like the place and length of employment, you might add something like, *"effectively implemented and communicated daily workflow to fellow team members, resulting in an increase in productivity."*

we can also use a statement to point out teamwork as an important quality to bring to the data analyst world. While you might have plenty of work to do on your own, it'll always be for the

benefit of the team. Team means not only the data team you're part of, but the whole company as well.

Soft skills ✧.*
non-technical traits and behaviors that relate to how you work. Being detail-oriented and demonstrating perseverance are two more examples of soft skills that anyone hiring a data analyst will look for.

Adding professional skills on your resume

1. Structured Query Language (SQL): SQL is considered a basic skill that is pivotal to any entry-level data analyst position. SQL helps you communicate with databases, and more specifically, it is designed to help you retrieve information from databases. Every month, thousands of data analyst jobs posted require SQL, and knowing how to use SQL remains one of the most common job functions of a data analyst.
2. Spreadsheets: Although SQL is popular, 62% of companies still prefer to use spreadsheets for their data insights. When getting your first job as a data analyst, the first version of your database might be in spreadsheet form, which is still a powerful tool for reporting or even presenting data sets. So, it is important for you to be familiar with using spreadsheets for your data insights.
3. Data visualization tools: Data visualization tools help to simplify complex data and enable the data to be visually understood. After gathering and analyzing data, data analysts are tasked with presenting their findings and making that information simple to grasp. Common tools that are used in data analysis include Tableau, Microstrategy, Data Studio, Looker, Datarama, Microsoft Power BI, and many more. Among these, Tableau is best known for its ease of use, so it is a must-have for beginner data analysts. Also, studies show that data analysis jobs requiring Tableau are expected to grow about 34.9% over the next decade.
4. **R or Python programming:** Since only less than a third of entry-level data analyst positions require knowledge of Python or R, you don’t need to be proficient in programming languages as an entry-level data analyst. But, R or Python are great additions to have as you become more advanced in your career.

Adding soft skills on your resume

1. **Presentation skills**
Although gathering and analyzing data is a big part of the job, presenting your findings in a clear and simple way is just as important. You will want to structure your findings in a way that allows your audience to know exactly what conclusions they are supposed to draw.
2. **Collaboration**

As a data analyst, you will be asked to work with lots of teams and stakeholders—sometimes internal or external—and your ability to share ideas, insights, and criticisms will be crucial. It is important that you and your team—which might consist of engineers and researchers—do your best to get the job done.

3. **Communication**

Data analysts must communicate effectively to obtain the data that they need. It is also important that you are able to work and clearly communicate with teams and business leaders in a language that they understand.

4. **Research**

As a data analyst, even if you have all of the data at your disposal, you still need to analyze it and draw crucial insights from it. To analyze the data and draw conclusions, you will need to conduct research to stay in-line with industry trends.

5. **Problem-solving skills**

Problem-solving is a big part of a data analyst's job, and you will encounter times when there are errors in databases, code, or even the capturing of data. You will have to adapt and think outside the box to find alternative solutions to these problems.

6. **Adaptability**

In the ever-changing world of data, you have to be adaptable and flexible. As a data analyst, you will be working across multiple teams with different levels of needs and knowledge, which requires you to adjust to different teams, knowledge levels, and stakeholders.

7. **Attention to detail**

A single line of incorrect code can throw everything off, so paying attention to detail is critical for a data analyst. When it comes to understanding and reporting findings, it helps if you focus on the details that matter to your audience.

Adding soft skills to your resume

Here are a few ways that you can add soft skills to your resume:

1. Analyze your previous work experience and find opportunities to insert a soft skill. For example, if you worked in a restaurant, you could emphasize your communication and adaptability skills that you utilized to effectively function during peak hours.
2. Call attention to your problem-solving, presentation, research, and communication skills in previous projects or relevant coursework.
3. Add a mix of soft and professional skills in the skills or summary section of your resume.

Where does your interest lie?

Junior or associate data analysts

Healthcare analyst ✧.*

They gather and interpret data from sources like electronic health records and patient surveys. Their work helps organizations improve the quality of their care. Health care analysts might also look for ways to lower the cost of care and improve patient experience.

Marketing analyst ✧.*

Data analysts in marketing complete quantitative and qualitative market analysis. They identify important statistics and interpret and present their findings to help stakeholders understand the data behind their marketing strategies.

Business intelligence analyst ✧.*

They help companies use data they've collected to increase their efficiency and maximize their profits. These analysts usually work with large amounts of data to identify trends and generate business insights.

Financial analyst ✧.*

use the data to identify and potentially recommend business and investment opportunities**.** if you're a junior analyst in this field, you'll probably start off doing a lot of data gathering and financial modeling as well as spreadsheet maintenance.