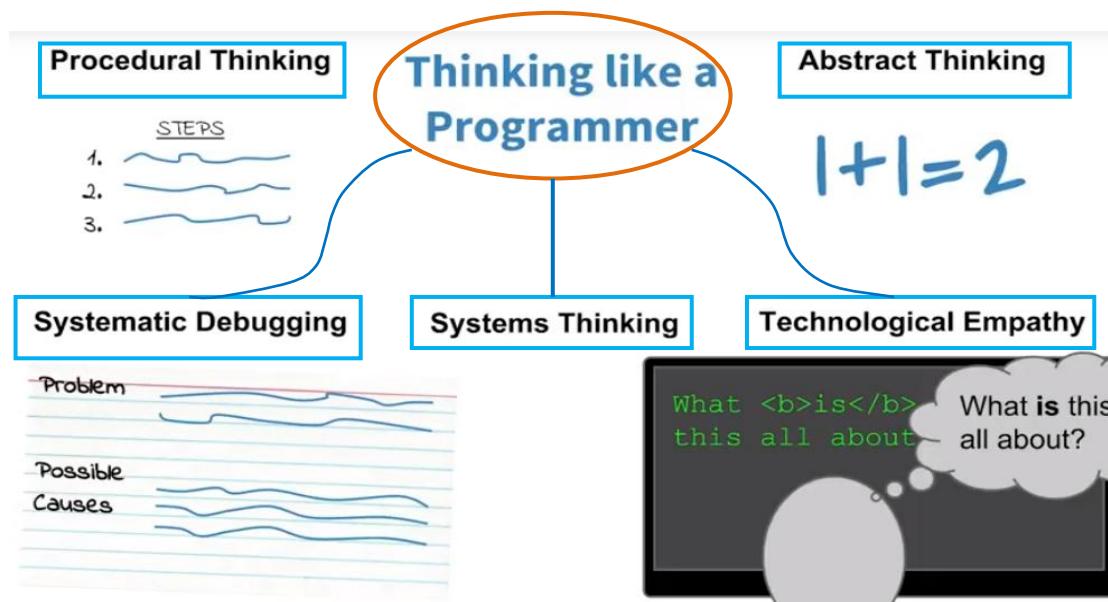


LESSON 1

Thinking Like a Programmer

We'll focus on 5 way to think :

1. Procedural thinking, and this involves taking a big problem and breaking the solution into set of step.
2. Abstract thinking. And if you couldn't do abstract thinking, you may know that if someone gave you one apple and then another, you have two. But if someone asked you what happened when they gave one orange and then another orange, you might have no idea. But because you do have an understanding of the abstract idea of numbers, you know that you have two oranges. But even more importantly, you know the abstract idea, that one plus one is two.
3. A Systematic way of solving problems, or systematic debugging. This is the skill where you can solve systematically problems, or potential problems, and then track down potential causes, until you find the problem and fix it. (method logical method for solving errors)
4. Technological empathy, and it involves understanding how the computer sees thing, and you've already started developing this. Now, when you see texts like this, you can start to think about it the way a computer would. You may already be thinking that the "is" should be bold.
5. Systems thinking. This may sound a little intimidating, but basically this involves taking one big idea, like, what does it mean to think like a programmer. And break it down into its component pieces.



Thinking Like a Programmer

Thinking Like a Programmer

A large goal of this program is to give you practice *thinking* like a computer programmer. Longtime programmers will often talk about how the programming thought process has helped them in other areas of their life, but what does it mean to "think like a programmer?" In this program we will focus on five specific ways of thinking that all programmers have to use all the time. .

1. Procedural Thinking

Procedural thinking involves creating perfectly clear and unambiguous instructions for a computer to follow.

It may be easy to tell a person how to sort a deck of cards, but getting a computer to do that requires more thought. But once you've learned to think this way you'll find that computers can do it **much** faster.

2. Abstract Thinking

Abstract thinking means finding similarity, or as programmers would say, **generality** amongst seemingly different things.

In this program you will visit hundreds of Udacity web pages. It would be impractical for a programmer to specifically program each of these pages individually. The Udacity programmers used the power of what's called *abstraction* to avoid unnecessary repetition of work.

3. Systems Thinking

Systems thinking happens when you break a big problem down into smaller pieces. Programmers do this when they create a plan (often on paper) for how a program will work. It involves big-picture thinking and decision-making about a problem and how different pieces of a program can work together to solve it. If this is unclear now, don't worry!

4. Technological Empathy

"Technological empathy" comes in many forms. For example, computer empathy is the ability to understand what a computer is, how it works, and what it's good and bad at doing.

A computer is a tool, as are the programs we use and the programming languages used to write them. It's almost impossible to program anything substantial without a basic understanding of how these tools work.

5. Debugging

Debugging is a systematic process of relentlessly identifying the cause of a computer program that doesn't work.

When a program doesn't work, it's because there is a mistake (also known as a "bug") somewhere in the computer code. Since these mistakes are an unavoidable part of programming, every good programmer has a system for fixing them and that system usually looks something like this:

- 1.** Collect evidence (what makes this program fail?)
- 2.** Generate theories (what may have caused this problem?)
- 3.** Test those theories (if my theory is correct, how could I find out?)
- 4.** Fix the problem

Learning Mindset

Thinking Like a Programmer

A large goal of this program is to give you practice thinking like a computer programmer. Longtime programmers will often talk about how the programming thought process has helped them in other areas of their life, but what does it mean to "think like a programmer?" In this program we will focus on five specific ways of thinking that all programmers have to use all the time. .

1. Procedural Thinking

Procedural thinking involves creating perfectly clear and unambiguous instructions for a computer to follow.

It may be easy to tell a person how to sort a deck of cards, but getting a computer to do that requires more thought. But once you've learned to think this way you'll find that computers can do it **much** faster.

2. Abstract Thinking

Abstract thinking means finding similarity, or as programmers would say, **generality** amongst seemingly different things.

3. Systems Thinking

Systems thinking happens when you break a big problem down into smaller pieces. Programmers do this when they create a plan (often on paper) for how a program will work. It involves big-picture thinking and decision-making about a problem and how different pieces of a program can work together to solve it.

4. Technological Empathy

"Technological empathy" comes in many forms. For example, computer empathy is the ability to understand what a computer is, how it works, and what it's good and bad at doing.

A computer is a tool, as are the programs we use and the programming languages used to write them. It's almost impossible to program anything substantial without a basic understanding of how these tools work.

5. Debugging

Debugging is a systematic process of relentlessly identifying the cause of a computer program that doesn't work.

When a program doesn't work, it's because there is a mistake (also known as a "bug") somewhere in the computer code. Since these mistakes are an unavoidable part of programming, **every good programmer has a system for fixing them and that system usually looks something like this:**

1. Collect evidence (what makes this program fail?)
2. Generate theories (what may have caused this problem?)
3. Test those theories (if my theory is correct, how could I find out?)
4. Fix the problem

Learning Mindset

To be successful, you'll need to be self-motivated to learn. This program will be challenging, and it may be tempting to take a long break or even stop when you get stuck. Students who persevere, are consistent, and seek out answers on their own are best suited to succeed. Keep asking questions, and never give up! In a later lesson, we'll go into more detail on the topic of learning mindsets.

Skills to Get Started

For now, there are some basic computer skills that will be vital to success. Take a look at each of the skills listed below, and if you're not fully comfortable with any of them, scroll down to the Resources and Tutorials section that follows to find more in-depth instructions. You may want to bookmark this page for future reference.

Basic computer skills you'll need:

- Finding the location of a file on your computer
- Renaming a file
- Compressing and uncompressing a file (using a zip file)
- Downloading and installing programs from the Internet

Resources and Tutorials

Below is a collection of resources and tutorials to familiarize you with each of the computer skills listed above that you'll need to be successful in this program.

Finding the location of a file on your computer

Windows — finding a file or folder

Mac OS X (Finder)

Mac OS X (paths)

Renaming a file on your computer

Mac OS X

Windows — renaming a file

Compressing and uncompressing a file (using a zip file)

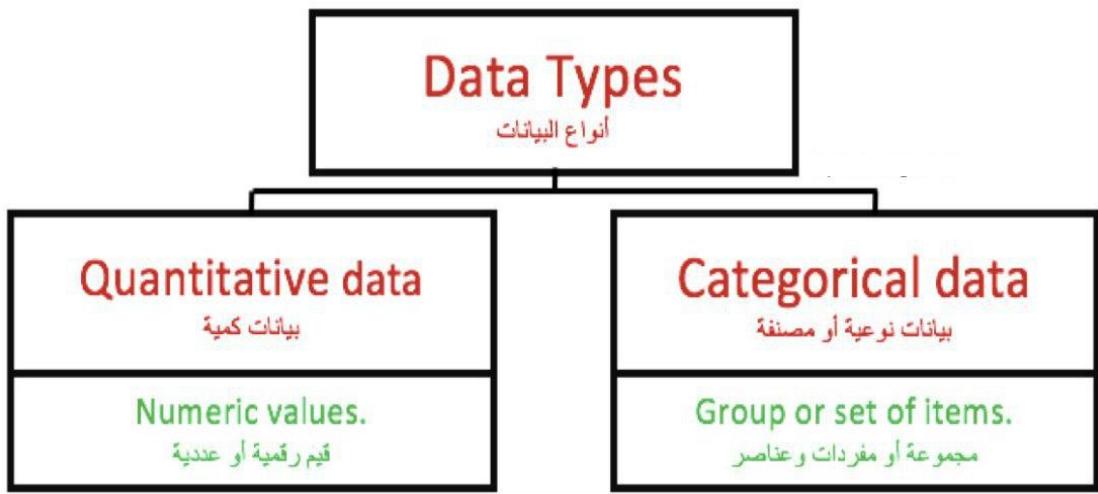
Mac OS X

Windows

Installing programs downloaded from the Internet to your computer

Mac — installing a program. **Note:** Once you've installed a program, you can eject the installation file with the keyboard shortcut: **Command + e**

Windows — installing a program. **Note:** Once you've installed a program, you can delete the installation file.



Example: Weight by kg, Number of factories in a city.
مثال: الوزن بالكيلو جرام، عدد المصانع في مدينة ما.

Example: Gender (male or female), Letter grades (A+, A, B+, B, ...).
مثال: الجنس (ذكر أو أنثى)، تقديرات النجاح (أ+, أ, ب+, ب, ...)

Categorical Data Types

Categorical Ordinal vs. Categorical Nominal :

We can divide categorical data further into two types: **Ordinal** and **Nominal**.

Categorical Ordinal data take on a ranked ordering (like a ranked interaction on a scale from **Very Poor** to **Very Good** with the dogs).

Categorical Nominal data do not have an order or ranking (like the breeds of the dog).

CATEGORICAL DATA TYPES	
ORDINAL (ordered)	NOMINAL (No order)
Rating	Breed
Very Positive	Lab
Positive	Boodle
Neutral	PUG
Negative	Chihuahua
Very Negative	Greyhound

Quantitative Data Types

Continuous vs. Discrete

We can think of quantitative data as being either **continuous** or **discrete**.

Continuous data can be split into smaller and smaller units, and still a smaller unit exists. An example of this is the age of the dog - we can measure the units of the age in years, months, days, hours, seconds, but there are still smaller units that could be associated with the age.

Discrete data only takes on countable values. The number of dogs we interact with is an example of a discrete data type.

Note : the Continuous values can take the form of any numeric value, including decimal values, and sometimes take negative numbers.



Quantitative Data Types		
	Continuous Data	Discrete Data
Example	Age of dog	Number of dogs
Continuous values that can be Split into smaller values.		
Continuous values that are Countable.		
Continuous values can take the form of any numerical value including Decimal values and sometimes even negative numbers.	Discrete values only take countable values.	

Recap

The table below summarizes our data types. To expand on the information in the table, you can look through the text that follows.

Data Types

Quantitative:	Continuous	Discrete
	Height, Age, Income	Pages in a Book, Trees in Yard, Dogs at a Coffee Shop
Categorical:	Ordinal	Nominal
	Letter Grade, Survey Rating	Gender, Marital Status, Breakfast Items

Below is a little more detail of the information shared in the above table.

You should have now mastered what types of data in the world around us falls into each of these four buckets: Discrete, Continuous, Nominal, and Ordinal. In the next sections, we will work through the numeric summaries that relate specifically to quantitative variables.

Quantitative Vs. Categorical

Some of these can be a bit tricky - notice even though zip codes are a number, they aren't really a quantitative variable. If we add two zip codes together, we do not obtain any useful information from this new value. Therefore, this is a categorical variable.

Height, Age, the **Number of Pages in a Book** and **Annual Income** all take on values that we can add, subtract and perform other operations with to gain useful insight. Hence, these are **quantitative**.

Gender, Letter Grade, Breakfast Type, Marital Status, and **Zip Code** can be thought of as labels for a group of items or individuals. Hence, these are **categorical**.

Continuous Vs. Discrete

To consider if we have continuous or discrete data, we should see if we can split our data into smaller and smaller units. Consider time - we could measure an event in years, months, days, hours, minutes, or seconds, and even at seconds we know there are smaller units we could measure time in. Therefore, we know this data type is continuous. **Height, age**, and **income** are all examples of **continuous data**. Alternatively, the **number of pages in a book**, **dogs I count outside a coffee shop**, or **trees in a yard** are **discrete data**. We would not want to split our dogs in half.

Ordinal Vs. Nominal

In looking at categorical variables, we found **Gender, Marital Status, Zip Code** and your **Breakfast items** are **nominal variables** where there is no order ranking associated with this type of data. Whether you ate cereal, toast, eggs, or only coffee for breakfast; there is no rank ordering associated with your breakfast.

Alternatively, the **Letter Grade** or **Survey Ratings** have a rank ordering associated with it, as **ordinal data**. If you receive an A, this is higher than an A-. An A- is ranked higher than a B+, and so on... Ordinal variables frequently occur on rating scales from very poor to very good. In many cases we turn these ordinal variables into numbers, as we can more easily analyze them, but more on this later!

QUIZ QUESTION

This quiz will assure you have a clear understanding of the differences between categorical nominal vs. categorical ordinal variables. All of the below variables are categorical. It is your job to **check** all of the below which are **nominal**. Do not check the ordinal categorical variables.

- Letter Grades (A, B+, B, B-, etc.)
- Types of Fruit (Apple, Banana, etc.)

- Ratings on a Survey (Poor, Ok, Great)
- Types of Dog Breeds (German Shepherd, Collie, etc.)

- Genres of Movies (Horror, Comedy, etc.)

- Gender

- Nationality

- Education (HS, Associates, Bachelors, Masters, PhD, etc.)

This quiz will assure you have a clear understanding of the differences between quantitative continuous vs. discrete variables. All of the below variables are quantitative. It is your job to **check** all of the below which are **continuous**. Do not check the discrete variables.

Travel Distance from Home to Work

Number of Pages in a Book

Amount of Rain in a Year

Time to Run a Mile

Number of Movies Watched in a Week

Amount of Water Consumed in a Day

Number of Phones per Household

Measures of Center (Mean)

Analyzing Quantitative Data

Four Aspects for Quantitative Data

There are four main aspects to analyzing **Quantitative** data.

1. Measures of **Center**
2. Measures of **Spread**
3. The **Shape** of the data.
4. **Outliers**

Analyzing Categorical Data

Though not discussed in the video, analyzing categorical data has fewer parts to consider. **Categorical** data is analyzed usually by looking at the counts or proportion of individuals that fall into each group. For example if we were looking at the breeds of the dogs, we would care about how many dogs are of each breed, or what proportion of dogs are of each breed type.

Measures of Center

There are three measures of center:

1. **Mean**
2. **Median**
3. **Mode**

The Mean

In this video, we focused on the calculation of the mean. The mean is often called the average or the **expected value** in mathematics. We calculate the mean by adding all of our values together, and dividing by the number of values in our dataset.

The remaining measures of the median and mode will be discussed in detail in the upcoming quizzes and lectures.

To illustrate how each of these measurements is calculated, we will take the following example:

This table is specific to the number of dogs seen at the coffee shop during a week

MON	TUE	WED	THU	FRI	SAT	SUN
5	3	8	3	15	45	9

Look at the table

How many dogs are expected to see on a particular day?

The answer to this question is chosen in different ways, For example, depending on the day or the week.

The expected answer is usually associated with the mean or average of our data set.

How to calculate the mean:

Sum of all values divided by the count of values

$$\frac{5+3+8+3+15+45+9}{7} = 12.57 \text{ dogs}$$

Note :

The center is not always the best measure of the center.

For the data set in the table it can be seen that the mean does not really look like it is in the middle of the data at all.

In just two days of 7 days the number of dogs was recorded greater than the mean for them.

The center divides our dogs into decimal values that will look strange when answering our question.

Quiz : Measures of Center (Mean)

QUESTION 1 OF 2

Which of the below are measures of center (Check all that apply)?

Mean

Standard Deviation

Variance

Median

Inter-quartile Range

Mode

Range

Maximum

Minimum

QUESTION 2 OF 2

If we have the data:

5, 8, 15, 7, 10, 22, 3, 1, 15

What is the mean?

Solution:

$$\text{Mean} = \frac{\sum_{i=1}^n x_i}{n}$$

7

9.56

15

8

8.5

Measures of Center (Median)

The Median

The **median** splits our data so that 50% of our values are lower and 50% are higher. We found in this lecture that how we calculate the median depends on if we have an even number of observations or an odd number of observations.

Example:

For the data set in the previous example

MON	TUE	WED	THU	FRI	SAT	SUN
5	3	8	3	15	45	9

We will sort the values in the table ascending or descending

3 3 5 **8** 9 15 45

We have a broker equal to 8 and this is a better answer than a 12.57 dog that is shown by the middle, and not only 8 in the middle of the data set. But does not divide any of the dogs into halves.

As a note about the median account, the actual account of the broker depends on our dealings with a data set with an even number or odd number of data.

Median for Odd Values

If we have an **odd** number of observations, the **median** is simply the number in the **direct middle**. For example, if we have 7 observations, the median is the fourth value when our numbers are ordered from smallest to largest. If we have 9 observations, the median is the fifth value.

Median for Even Values

If we have an **even** number of observations, the **median** is the **average of the two values in the middle**. For example, if we have 8 observations, we average the fourth and fifth values together when our numbers are ordered from smallest to largest.

In order to compute the median we MUST sort our values first.

Let us give two examples of clarification

i. $1, 2, 3, 3, 5, 8, 10 \quad n = 7 \Rightarrow \text{odd}$

In this case we deal with seven values. In the coding we can say $n = 7$, and because it is an odd number, we have the median exactly like the mean value, in which case the median is 3.

ii. $5, 8, 3, 2, 1, 3, 10, 105 \quad n = 8 \Rightarrow \text{even}$

In this case it is considered to be $n= 8$ (even number).

First we have to arrange these values,

$$\textcolor{blue}{1, 2, 3, 3, 5, 8, 10, 105}$$

Since there is no accurate central value, in this case we will take the average of the two values in the center as the median.

Note, here does not require the number to be in our data set, because n is equal to 8 In this example, the average of the fourth and fifth values shows us the argument for this dataset

$$\frac{\textcolor{blue}{3+5}}{2} = \textcolor{blue}{4}$$

Whether we use the mean or median to describe a dataset is largely dependent on the **shape** of our dataset and if there are any **outliers**. We will talk about this in just a bit!

Quiz : Measures of Center (Median)

QUESTION 1 OF 2

If we have the data:

5, 8, 15, 7, 10, 22, 3, 1, 15

What is the median?

7

9.56

15

8

7.5

QUESTION 2 OF 2

If we have the data:

5, 8, 15, 7, 10, 22, 3, 1, 15, 2

What is the median?

7

9.56

15

8

7.5

Measures of Center (Mode)

The Mode

The **mode** is the most frequently observed value in our dataset.

In another ward, the mode is the most frequent number in a dataset.

There might be multiple modes for a particular dataset, or no mode at all.

Third measure aims to provide us with the most common value in this data set

1, 2, **3**, **3**, 5, 8, 10

This value is equal to **3** and is called the mode.

No Mode

If all observations in our dataset are observed with the same frequency, there is no mode. If we have the dataset:

1, 1, 2, 2, 3, 3, 4, 4

There is no mode, because all observations occur the same number of times.

Many Modes

If two (or more) numbers share the maximum value, then there is more than one mode.

If we have the dataset:

1, 2, 3, 3, 3, 4, 5, 6, 6, 6, 7, 8, 9

There are two modes 3 and 6, because these values share the maximum frequencies at 3 times, while all other values only appear once.

Quiz : Measures of Center (Mode)

QUESTION 1 OF 4

Check all of the below that are true with regards to our measures of center.

- The mode is the middle number in the dataset when the numbers are rank ordered.

- The median is the middle number in the dataset when the numbers are rank ordered.

- The mean is always the best measure of center for any dataset.

- The mean is always less than the median.

- The median is always the best measure of center for any dataset.

- The mode is always the best measure of center for any dataset.

QUESTION 2 OF 4

If we have the data:

5, 8, 15, 7, 10, 22, 3, 1, 15

What is the mode?

7

9.56

9

15

5

QUESTION 3 OF 4

For the dataset below match the correct measure to the value:

8, 12, 32, 10, 3, 4, 4, 4, 4, 5, 12, 20

Submit to check your answer choices!

MEASURE

VALUE

Mean

9.83

Median

6.5

Mode

4

None of the Above

8

QUESTION 4 OF 4

If we have the data:

5, 8, 15, 7, 10, 22, 3, 1, 15, 10

Mark all statements that are true.

The mode is 15.

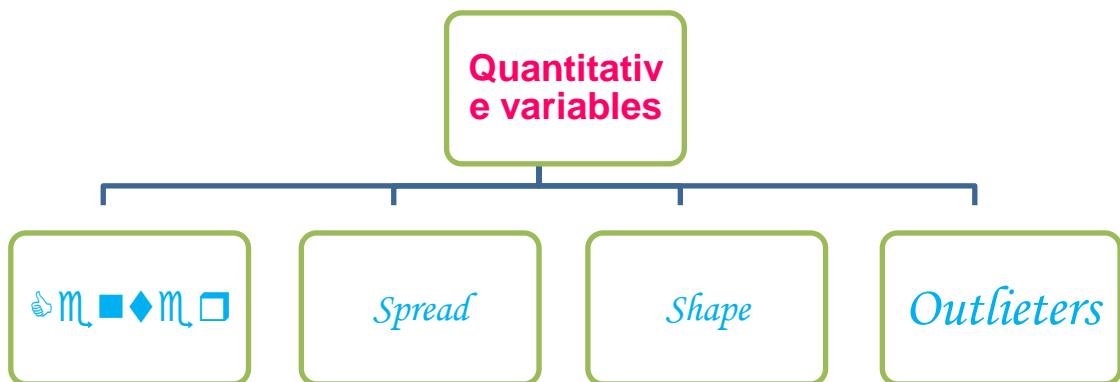
The mean is 15.

The mode is 10.

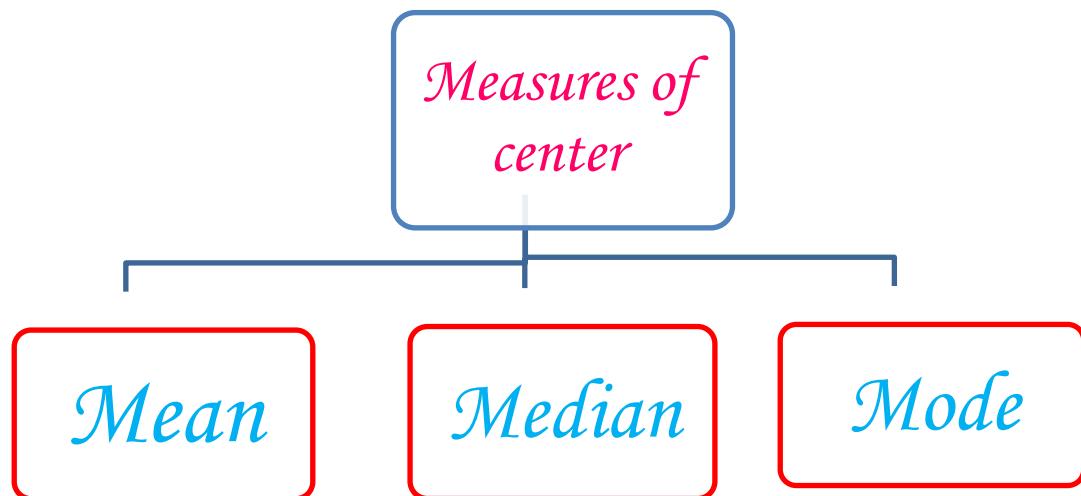
None of the above are true.

What is Notation?

Previously we mentioned the four basic aspects of quantitative data analysis.



We also looked specifically at



Note : Before we look at spreading the values important to understand the notation.

- The notation is a common mathematical language used to communicate.
- Coding is a necessary tool for communicating ideas associated with data.

Notation

Notation is a common language used to communicate mathematical ideas. **Think of notation as a universal language used by academic and industry professionals to convey mathematical ideas.**

You likely already know some notation. Plus, minus, multiply, division, and equal signs all have mathematical symbols that you are likely familiar with. Each of these symbols replaces an idea for how numbers interact with one another. In the coming concepts, you will be introduced to some additional ideas related to notation. Though you will not need to use notation to complete the project, **it does have the following properties:**

1. **Understanding how to correctly use notation makes you seem really smart.** Knowing how to read and write in notation is like learning a new language. A language that is used to convey ideas associated with mathematics.
2. **It allows you to read documentation, and implement an idea to your own problem.** Notation is used to convey how problems are solved all the time. One really popular mathematical algorithm that is used to solve some of the world's most difficult problems is known as Gradient Boosting. The way that it solves problems is explained here: https://en.wikipedia.org/wiki/Gradient_boosting. If you really want to understand how this algorithm works, you need to be able to read and understand notation.
3. **It makes ideas that are hard to say in words easier to convey.** Sometimes we just don't have the right words to say. For those situations, I prefer to use notation to convey the message. Similar to the way an emoji or meme might convey a feeling better than words, notation can convey an idea better than words. Usually those ideas are related to mathematics.

Random Variables

Rows and Columns

Spreadsheets are a common way to hold data. They are composed of rows and columns. Rows run horizontally, while columns run vertically. Each column in a spreadsheet commonly holds a specific **variable**, while each row is commonly called an **instance** or **individual**.

The example in the shown below.

Date	Day of Week	Time Spent On Site (X)	Buy (Y)
June 15	Thursday	5	No
June 15	Thursday	10	Yes
June 16	Friday	20	Yes

This is a row:

Date	Day of Week	Time Spent On Site (X)	Buy (Y)
June 15	Thursday	5	No

This is a column:

Time Spent On Site (X)
5
10
20

Before collecting data, we usually start with a question, or many questions, that we would like to answer. The purpose of data is to help us in answering these questions.

Random Variables

A **random variable** is a placeholder for the possible values of some process (mostly... the term 'some process' is a bit ambiguous). As was stated before, notation is useful in that it helps us take complex ideas and simplify (often to a single letter or single symbol). We see random variables represented by capital letters (**X**, **Y**, or **Z** are common ways to represent a random variable).

We might have the random variable **X**, which is a holder for the possible values of the amount of time someone spends on our site. Or the random variable **Y**, which is a holder for the possible values of whether or not an individual purchases a product.

X is 'a holder' of the values that could possibly occur for the amount of time spent on our website. Any number from 0 to infinity really.

To understand how to use data tables for data retention, we will take an example.

Suppose that someone runs a small blog about :

- His best adventures with dogs and the worst in the cafe.
- The coffee shop sells the jewels related to these adventures and everything from the throwing games to the leashes and to the bags of dogs and everything between them.

Of the questions we have:

- How many people visit the site?
- How much time do visitors spend on the site?
- Are there differences in traffic depending on the day of the week?
- How many visitors purchase an item through the blog?

Solution :

To answer the questions, we assume that we are going after the date of the visit, the day your week, the amount of time spent on the site, whether someone will buy an element or not.

DATE	DOW	TIME	BUY
Jun 15	Thur	5	NO
Jun 15	Thur	10	Yes
Jun 16	Fri	7	Yes
Jun 16	Fri	9	Yes
Jun 16	Fri	12	No

- A column in our dataset is associated with a random variable.
- Explaining the concept of a random variable in English or any other language is complex, but in notation, it is simple.

To answer the question of the time visitors take on the site we need to take a look at the time column.

To answer the question: Are there differences in the traffic ratio by day of the week we will look at the Dow column.

To answer the question of how many purchases are made through the blog, take a look at the Buy column.

Mathematically, we usually assume a column or a random variable using a large letter, and usually we use an X character, but we can use a Y, Z or any other big character.

- ⊕ Assume that the random variable X which indicates the amount of time an individual takes on the site is then associated with the X elements of the rest of the entire column.
- ⊕ Assume that the random variable w_i which indicates whether an individual will buy an element from the site or not is then associated with any other column elements.

Similarly, we follow the same method of naming variables or columns in a table.

Quiz : Variable Types

Example Dataset

An example of the data we might have collected in the previous video is shown here:

Date	Day of Week	Time Spent On Site (X)	Buy (Y)
June 15	Thursday	5	No
June 15	Thursday	10	Yes
June 16	Friday	20	Yes

QUESTION 1 OF 2

What type of variable is the random variable X in the video in the previous concept?

- Categorical - Ordinal
- Categorical - Nominal
- Quantitative - Continuous
- Quantitative - Discrete

QUESTION 2 OF 2

What type of variable is the random variable Y in the video in the previous concept?

- Categorical - Ordinal
- Categorical - Nominal
- Quantitative - Continuous
- Quantitative - Discrete

Capital vs. Lower

Capital vs. Lower Case Letters

Random variables are represented by capital letters. Once we observe an outcome of these random variables, we note it as a lower case of the same letter.

Example 1

Taking this one step further, we could ask:

What is the probability someone spends more than 20 minutes in our website?

In notation, we would write:

P(X > 20)?

Here **P** stands for **probability**, while the parentheses encompass the statement for which we would like to find the probability. Since **X** represents the amount of time spent on the website, this notation represents the probability the amount of time on the website is greater than 20.

We could find this in the above example by noticing that only one of the 5 observations exceeds 20. So, we would say there is a **1 (the 45) in 5 or 20%** chance that an individual spends more than 20 minutes on our website (based on this dataset).

Example 2

If we asked:

What is the probability of an individual spending 20 or more minutes on our website?

We could notate this as:

$$P(X \geq 20)?$$

We could then find this by noticing there are two out of the five individuals that spent 20 or more minutes on the website. So this probability is **2 out of 5 or 40%**.

Based on what we learned in the previous lesson we note that:

1. The X is not a number, it is a full set of possible values.
2. We can think of the X character as a placeholder for any possible values.
3. When we look at one result of the random variable, we use a lowercase letter with a character to the specified value of the random variable we are thinking.
4. the notation is a necessary tool for us to convey mathematical ideas.

Quiz : Introduction to Notation

Consider we have the following table:

Years Experience	Department	Part/Full Time
5	IT	Part Time
10	Finance	Full Time
8	HR	Full Time
1	Finance	Part Time

Consider we have the following labels:

X= years of experience

Y= Department

Z= Part/Full Time

Match the following notation to their corresponding:

- A. x_1 B. y_2 C. z_3 D. n**

QUIZ QUESTION

Use the information above to match the correct notation label to its corresponding value.

Part Time Years Experience Department 16

NOTATION	VALUE
A. (this refers to the letter with the corresponding notation above)	5
B. (this refers to the letter with the corresponding notation above)	Finance
C. (this refers to the letter with the corresponding notation above)	Full Time
D. (this refers to the letter with the corresponding notation above)	4

Better Way?

In the following concepts, we will integrate what we know about how to calculate the mediator through coding.

Notation for Calculating the Mean

We know that the mean is calculated as the sum of all our values divided by the number of values in our dataset.

In our current notation, adding all of our values together can be extremely tedious. If we want to add 3 values of some random variable together, **we would use the notation:**

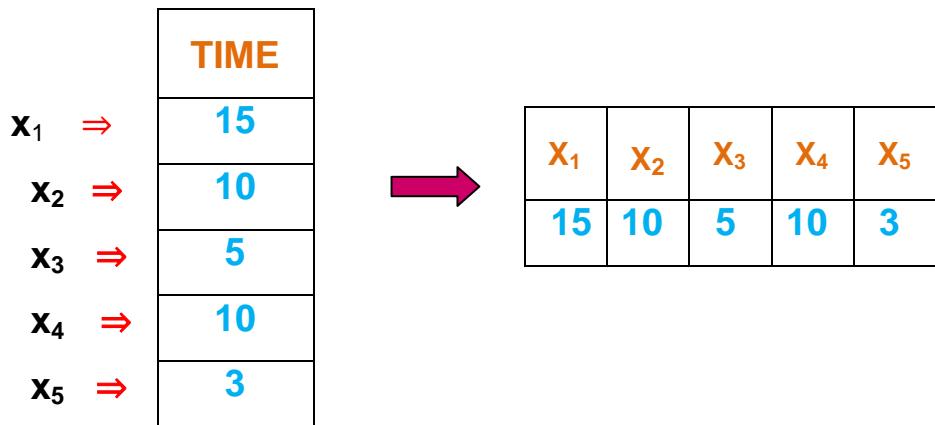
$$x_1 + x_2 + x_3 + \dots \dots$$

We will take example to illustrated :

Calculate how much time a person takes on the website in minutes. So that

X = Time spent on website "minutes"

Let's say we collect the amount of time that 5 people spend on the website.



If we want to collect more than two values, following the process itself will be boring, but it turns out that there is a better way.

Summation

Aggregations

An **aggregation** is a way to turn multiple numbers into fewer numbers (commonly one number).

Common assemblies include center measurements we have previously provided, such as mean, median, mode. Each of these assemblies takes numbers and provides a single value to provide information about the set of evidence.

Summation is a common aggregation. The notation used to sum our values is a Greek symbol called sigma Sigma Σ .

Example 1

Imagine we are looking at the amount of time individuals spend on our website. We collect data from nine individuals:

$$x_1 = 10, x_2 = 20, x_3 = 45, x_4 = 12, x_5 = 8, x_6 = 12, x_7 = 3, x_8 = 68, x_9 = 5$$

If we want to sum the **first three values** together in our previous notation, we write:

$$x_1 + x_2 + x_3$$

In our new notation, we can write:

$$\sum_{i=1}^3 x_i$$

Notice, our notation starts at the first observation ($i=1$) and ends at 3 (the number at the top of our summation). So all of the following are equal to one another:

$$x_1 + x_2 + x_3 = 10 + 20 + 45 = 75$$

Example 2

Now, imagine we want to sum the **last three values** together.

$$x_7 + x_8 + x_9$$

In our new notation, we can write:

$$\sum_{i=7}^9 x_i$$

Notice, our notation starts at the seventh observation ($i = 7$) and ends at 9 (the number at the top of our summation).

Other Aggregations

The Σ sign is used for aggregating using summation, but we might choose to aggregate in other ways. Summing is one of the most common ways to need to aggregate. However, we might need to aggregate in alternative ways. If we wanted to multiply all of our values together we would use a product sign Π , capital Greek letter pi. The way we aggregate continuous values is with something known as integration (a common technique in calculus), which uses the following symbol \int which is just a long s. We will not be using integrals or products for quizzes in this class, but you may see them in the future!

Notation for the Mean

Final Steps for Calculating the Mean

To finalize our calculation of the mean, we introduce n as the total number of values in our dataset. We can use this notation both at the top of our summation, as well as for the value that we divide by when calculating the mean.

$$\frac{1}{n} \sum_{i=1}^n x_i$$

Instead of writing out all of the above, we commonly write \bar{x} to represent the mean of a dataset.

Although, we could use any variable. Therefore, we might also write \bar{y} or any other letter.

We also could index using any other letter, not just i . We could just as easily use j, k or m to index each of our data values. The quizzes on the next concept will help reinforce this idea.

Notice

We should say $\sum_{i=1}^5 x_i = x_1 + x_2 + x_3 + x_4 + x_5$. The x_i is missing here in front of the summation.

Quiz : Summation

Match The Notation

For this quiz, you will be matching the notation attached the letters below to the corresponding numeric value to make sure you understand exactly what is being done with each part of the notation.

Imagine, we have the following table of values:

x_1	x_2	x_3	x_4	x_5	x_6	x_7
5	15	3	3	8	10	12

- A. n B. $\sum_{i=1}^n x_i$ C. $\sum_{j=2}^7 x_j + 6$ D. x_5 E. $\frac{\sum_{i=3}^6 x_i}{n-1}$

QUIZ QUESTION

Use the letters, numbers, and notation as defined above to match each letter to the appropriate value.

39 15

LETTER	VALUE
A.	7
B.	56
C.	57
D.	8
E.	4

Notation for Quizzes

For the below quiz, let the following letters denote the corresponding notation:

- A. X
- B. Y
- C. x_1
- D. n
- E. $\sum_{i=1}^n x_i$

QUESTION 1 OF 2

Use the letter next to the notation above to match the notation to the description of what the notation represents.

The notation for all of the values in our dataset.

The notation for all of the values in our dataset.

The notation for the number of columns in our dataset.

NOTATION LETTER

DESCRIPTION

A

the notation for a random variable.

B

the notation for a random variable.

C

the notation for the first observed Value of random variable.

D

the notation for the number of rows in our dataset.

E

the notation for the sum of all the values in our dataset.

Notation for Quizzes

For the below quiz, let the following letters denote the corresponding notation:

A. $\sum_{i=1}^n x_i$

B. $\frac{\sum_{i=1}^n x_i}{n}$

C. \bar{x}

D. \bar{y}

E. $\frac{\sum_{j=1}^n y_j}{n}$

QUESTION 2 OF 2

If we wanted to provide notation for the mean of a particular dataset, which of the following letters would correspond to the notation attached to calculating the mean? (Mark all that apply.)

A

B

C

D

E

Summary on Notation

Notation Recap

Notation is an essential tool for communicating mathematical ideas. We have introduced the fundamentals of notation in this lesson that will allow you to read, write, and communicate with others using your new skills!

Notation and Random Variables

As a quick recap, **capital letters** signify **random variables**. When we look at **individual instances** of a particular random variable, we identify these as **lowercase letters** with subscripts attach themselves to each specific observation.

For example, we might have **X** be the amount of time an individual spends on our website. Our first visitor arrives and spends 10 minutes on our website, and we would say \bold{x}_1 is 10 minutes.

We might imagine the random variables as columns in our dataset, while a particular value would be notated with the lower case letters.

Notation	English	Example
X	A random variable	Time spent on website
x_1	First observed value of the random variable X	15 mins
$\sum_{i=1}^n x_i$	Sum values beginning at the first observation and ending at the last	$5 + 2 + \dots + 3$
$\frac{1}{n} \sum_{i=1}^n x_i$	Sum values beginning at the first observation and ending at the last and divide by the number of observations (the mean)	$(5 + 2 + 3)/3$
\bar{x}	Exactly the same as the above - the mean of our data.	$(5 + 2 + 3)/3$

Notation for the Mean

We took our notation even farther by introducing the notation for summation \sum . Using this we were able to calculate the mean as:

$$\frac{1}{n} \sum_{i=1}^n x_i$$

In the next lesson, you will see this notation used to assist in your understanding of calculating various measures of spread. Notation can take time to fully grasp. Understanding notation not only helps in conveying mathematical ideas, but also in writing computer programs - if you decide you want to learn that too! Soon you will analyze data using spreadsheets. When that happens, many of these operations will be hidden by the functions you will be using. But until we get to spreadsheets, it is important to understand how mathematical ideas are commonly communicated. **This isn't easy, but you can do it!**

What are Measures of Spread?

We begin this lesson in search of the second aspect in terms of the analysis of quantitative variables and is the spread. When discussing proliferation measures, consider the numeric values associated with distance between points and each other. Another ward, "How far are points from one another ?"

Measures of Spread

Measures of Spread are used to provide us an idea of how spread out our data are from one another. Common measures of spread include:

1. [Range](#)
2. [Interquartile Range \(IQR\)](#)
3. [Standard Deviation](#)
4. [Variance](#)

Throughout this lesson you will learn how to calculate these, as well as why we would use one measure of spread over another.

Histograms

Histograms are super useful to understanding the different aspects of quantitative data.
the most common visual for quantitative data

In the upcoming concepts, you will see histograms used all the time to help you understand the four aspects we outlined earlier regarding a **quantitative variable**:

1. Center
2. Spread
3. Shape
4. Outliers

In this lecture we will take a look at how to create exact Histograms we can extract.

To understand how to create histograms we can extracted:

Suppose that we have following data set

1,2,2,4 5,7,8 9,12 15

First we need to divide our data into categories, to create the histogram we will choose how to split into categories in the end.

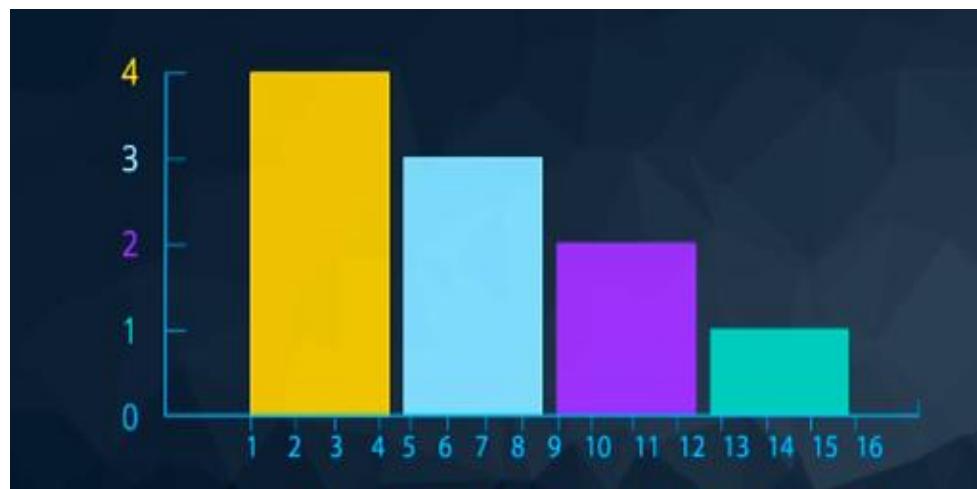
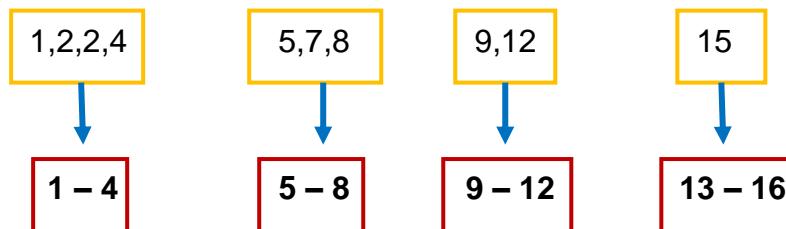
Here we have chosen our own categories

1 – 4 5 – 8 9 – 12 13 – 16

Because these values **1,2,2,4** fall between **1** and **4**, they move to the first category.

The values **5,7,8** are between **5** and **8**, so they fall into the following category.

The values **9** and **12** are in the following category and **15** are in the last category.

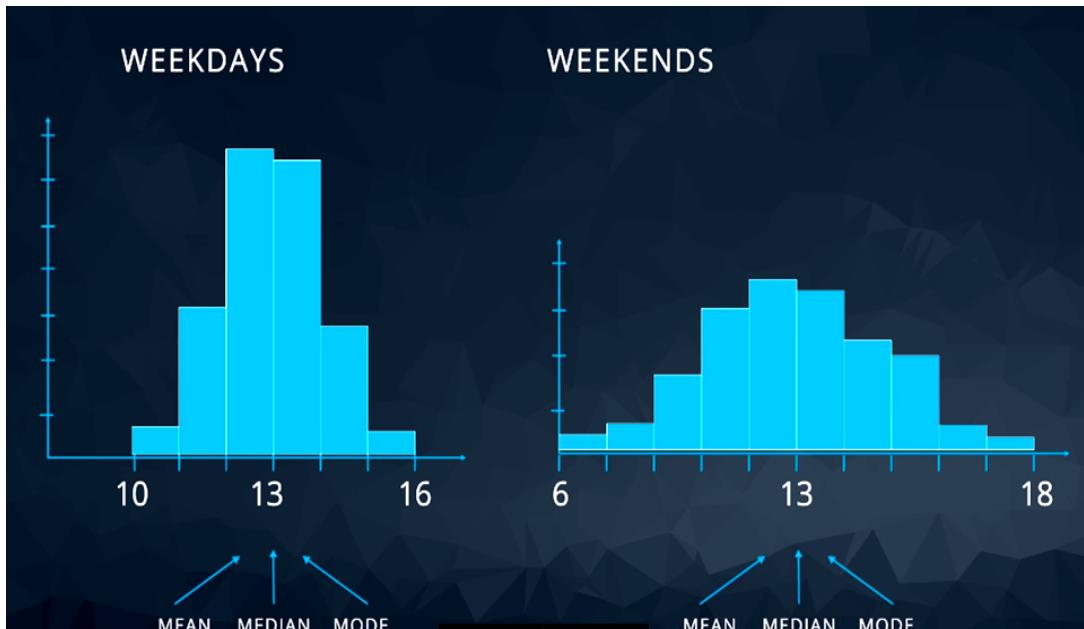


The number of values in each category determines the height of each bar on the histogram.

Changing categories will result in a slightly different view.

In fact, we do not have the right answer to choose our categories, and in most cases the program will choose the right categories for us, but it's something we should be aware of.

Weekdays vs. Weekends: What is the Difference



We have two histograms here to compare the number of dogs we saw on weekdays and the number of dogs we saw over the weekend of last year.

You will notice that the longest cells for each weekday and weekend were associated with **13** dogs, so the number of dogs that we expect to see is essentially the same on weekdays and weekends and the center's measures will basically be the same here.

Both histograms are mean, median and mode, and the three are equals about **13** dogs, but there is something different about these two distributions, what is the difference between them?

The difference is how to divide the data on each group. You can see the number of dogs you saw on weekdays, between **10** and **16**, while the number on weekends is between **6** and **18**.

In the next sections we will look at the most common ways to measure the spread of our data.

Introduction to Five Number Summary

Calculating the 5 Number Summary

One of the most common ways to measure the prevalence of our data is to look at summary number 5, which gives us the values to calculate the range and inter quartile range.

The five number summary consist of 5 values:

1. **Minimum:** The smallest number in the dataset.
2. **First quartile Q_1 :** The value such that **25%** of the data fall below.
3. **Second quartile "Median" Q_2 :** The value such that **50%** of the data fall below.
4. **Third quartile Q_3 :** The value such that **75%** of the data fall below.
5. **Maximum:** The largest value in the dataset.

In the above video we saw that calculating each of these values was essentially just finding the median of a bunch of different dataset. Because we are essentially calculating a bunch of medians, the calculation depends on whether we have an odd or even number of values.

Range

The **range** is then calculated as the difference between the **maximum** and the **minimum**.

IQR

The **interquartile range** is calculated as the difference between Q_3 and Q_1 .

In the upcoming sections, you will practice this with Katie and on your own.

Example (1):

Assume that we have the following odd data set :

$$5, 8, 3, 2, 1, 3, 10$$

Finding the 5 number summary.

Solution:

The first thing we need to calculate summary number 5 is to order the values in the data set.

$$\boxed{1, \boxed{2}, \boxed{3}, \boxed{3}, \boxed{5}, \boxed{8}, 10}$$

Once the values are arranged, it is easy to determine the largest and smallest values,

Where that **Maximum = 10 and Minimum = 1**.

While we calculate the median in the section on center measurements, we know that the median is the mean value in the data set **Median "Q₂" = 3**.

The remaining two values to complete the 5 number summary are Q₁ and Q₃. These values be thought of as the medians of the data on either side of Q₂.

$$Q_1 = 2 \quad \text{and} \quad Q_3 = 8$$

$$\text{Range} = \text{Maximum} - \text{Minimum} = 10 - 1 = 9$$

$$\text{Inter quartile range} = Q_3 - Q_1 = 8 - 2 = 6$$

Example (2):

Assume that we have the following even data set :

$$5, 8, 3, 2, 1, 3, 10, 105$$

Finding the 5 number summary.

Solution:

First, we arrange the values



$$\text{Maximum} = 10 \quad \text{and} \quad \text{Minimum} = 1$$

- When there is an even number of values, the median is equal to the mean of these **two values** $Q_2 = \frac{3+5}{2} = 4$

To find Q_1 and Q_3 divide our data set between the **two values** we are using to find a median.

$$Q_1 = \frac{2+3}{2} = 2.5 \quad \text{and} \quad Q_3 = \frac{8+10}{2} = 9$$

$$\text{Range} = \text{Maximum} - \text{Minimum} = 105 - 1 = 104$$

$$\text{Inter quartile range} = Q_3 - Q_1 = 9 - 2.5 = 6.5$$

Quiz : 5 Number Summary Practice

Do you know your 5 Number Summary?

Notice the dataset for the next 4 problems all use the same dataset.

What is the range associated with the following dataset?

1, 5, 10, 3, 8, 12, 4, 1, 2, 8

Solution:

1,1,2,3,4,5,8,8,10,12

Range = Max - Min = 12 - 1 = 11

Provided the values below, what is the value of the first quartile?

1, 5, 10, 3, 8, 12, 4, 1, 2, 8

Solution:

1, 1 , 2 , 3 , 4 , 5 , 8 , 8 , 10 , 12

$Q_1 = 2$

Provided the values below, what is the value of the third quartile?

1, 5, 10, 3, 8, 12, 4, 1, 2, 8

Solution:

1, 1 , 2 , 3 , 4 , 5 , 8 , 8 , 10 , 12

$Q_3 = 8$

Provided the values below, what is the value of the median? (this is the same as the second quartile)

1, 5, 10, 3, 8, 12, 4, 1, 2, 8

Solution:

$$1, 1, 2, 3, \boxed{4, 5}, 8, 8, 10, 12$$
$$Q_2 = \frac{4+5}{2} = 4.5$$

Do you know your 5 Number Summary?

Notice the dataset for the next 4 problems all use the same dataset.

Provided the values below, what is the value of the median? (this is the same as the second quartile)

5, 10, 3, 8, 12, 4, 1, 2, 8,

Solution :

$$1, 2, 3, 4, 5, 8, 8, 10, 12$$

$$Q_2 = 5$$

Provided the values below, what is the range? (The range should just be a single number)

5, 10, 3, 8, 12, 4, 1, 2, 8

Solution :

$$1, 2, 3, 4, 5, 8, 8, 10, 12$$

$$\text{Range} = 12 - 1 = 11$$

Provided the values below, what is the value for the first quartile?

5, 10, 3, 8, 12, 4, 1, 2, 8

Solution :

1, 2,3, 4, 5, 8, 8, 10, 12

$$Q_1 = \frac{2+3}{2} = 2.5$$

Provided the values below, what is the the value for the third quartile?

5, 10, 3, 8, 12, 4, 1, 2, 8

Solution :

1, 2, 3, 4, 5, 8, 8,10, 12

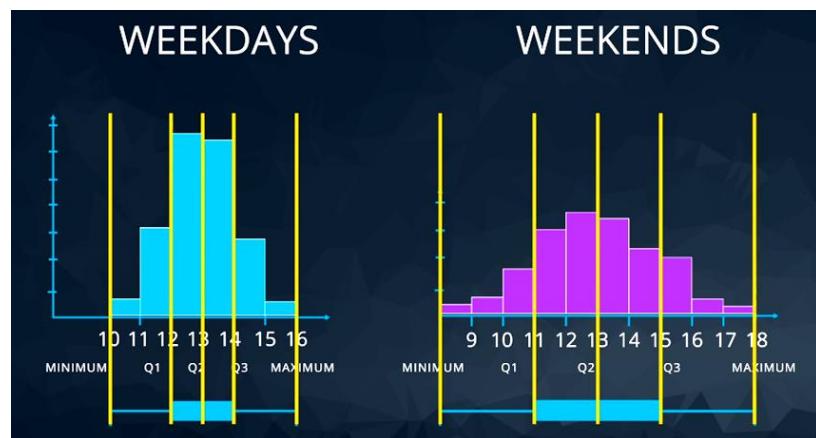
$$Q_3 = \frac{8+10}{2} = 9$$

What if We Only Want One Number?

Referring to the distributions we have found for the number of dogs we have seen, we can distinguish the values for number 5 in this format

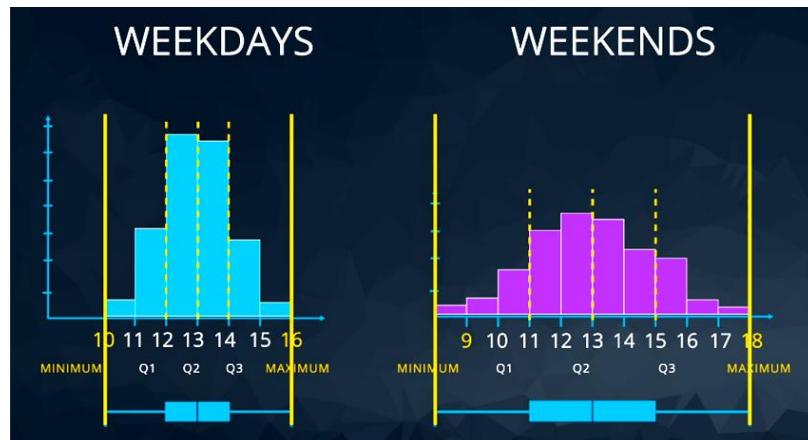


If we only take these marks, this is a common data plan known as the box Plot, although we prefer a repeatability hierarchy.

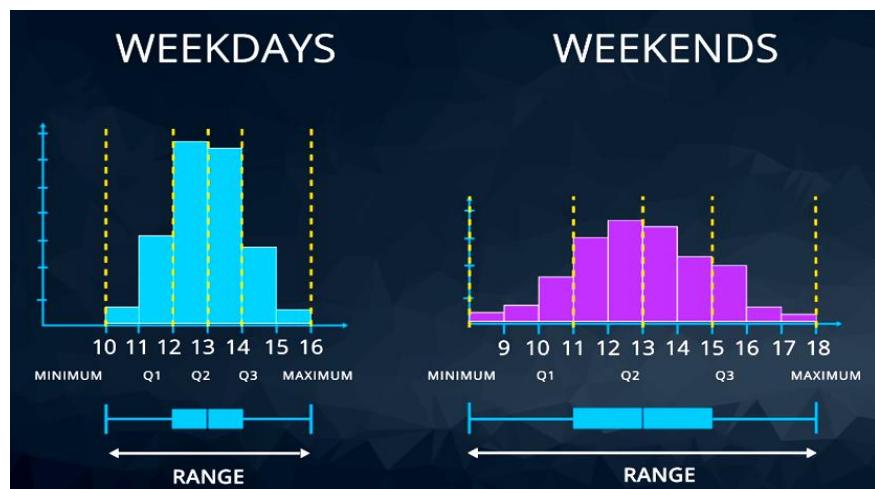


In most cases, the box plot can be useful for the speed of comparison between the spread of two sets of data across some key measurements, such as quadratic quadrants and maximum And minimum, from both histogram and summary number 5.

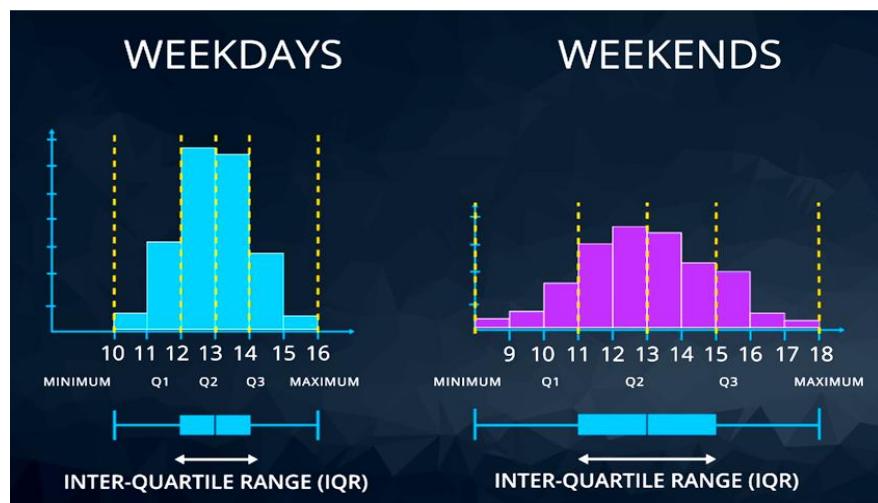
We can quickly see that the number of dogs we see on holiday is very different from the number of dogs we see on weekdays.



We can also visualize the distance from here to here as a range.



While the distance from here to here is a range between the two quarters.



There are many useful measurements that we can get from these Box Plot schemes. But what if we want to compare these distribution tables without having to take all 5 values for each distribution?

What if we want just one value to be able to compare the distribution schedules?

Introduction to Standard Deviation and Variance

Standard Deviation and Variance

The **standard deviation** is one of the most common measures for talking about the spread of data. It is defined as **the average distance of each observation from the mean**.

The most common method used by professionals to measure the spread of a data set using a single value is standard deviation or variation.

In this lesson We will focus on the standard deviation but in fact we will learn how to calculate the variance in the calculation.

When all's said and done with this calculation, the standard deviation will tell us on average. How far every point is from the mean of the points. As a quick mental picture, imagine we wanted to know how far employees were located from their place of work. One person might be 15 miles, another 35, another only one mile, and another might be remote and is 103 miles.

We could aggregate all of these distances together to show that the average distance employees are located from their work is 18 miles. But now, we want to know how the distance to work varies from one employee to the next.

We could use the 5 number summary as a description, but if we wanted just one number to talk about the spread, we'd probably choose the standard deviation.

For this example, this is how much on average the distance each person is from work differs from the average distance all of them are from work, so, this one is three miles farther from work than the average while this individual is four miles closer to work than the average.

The standard deviation is how far on average are individuals located from this mean distance. So, it's like the average of all these distances.

Standard Deviation Calculation

In the last lesson, we got an idea of what the standard deviation is measuring.

In this lesson, we will look at the math that actually occurs when calculating this measure. We will work with data to calculate this measure as well as associate notation with it. It's worth noting that after this lesson, you probably won't calculate this measure by hand ever again, because you'll learn software to do it for you.

Imagine we have a data set with four values, **(10, 14, 10, 6)**. The first thing we need to do to calculate the standard deviation is to find the mean.

$$\bar{x} = \frac{\sum_{i=1}^4 x_i}{n} = \frac{40}{4} = 10$$

In notation, we have this as **X - bar**. For our values, the sum is **40**, and we have **4** numbers. So, the mean is 40 over 4 or 10. Then we want to look at the distance of each observations from this mean. **$x_i - \bar{x} = 10 - 10 = 0, 14 - 10 = 4, 10 - 10 = 0, 6 - 10 = -4$** . Two of these observations are exactly equal to the mean. So, the distance here is zero, **($10 - 10 = 0, 10 - 10 = 0$)**. One is 4 larger than the 14, while the other is 4 smaller than the 6.

In notation, each of each these **$X_i - \bar{X}$** . then if we were to average these distances,

$(0 + 4 + 0 + (-4)) / 4 = 0$. the positive would cancel with the negative value. And the value of zero isn't a great measure of the spread here. Zero would suggest that all of these values are the same or that there's no spread. So instead, we need to make all of these values positive. The way we do this when calculating the standard deviation is by squaring them all.

$$(x_i - \bar{x})^2 = (10 - 10)^2 = 0^2 = 0, (14 - 10)^2 = 4^2 = 16, (10 - 10)^2 = 0^2 = 0, (6 - 10)^2 = (-4)^2 = 16$$

If we do that here, our negative and positive four values will become 16s .

Now, we could average these to find the average squared distance of each observation from the mean.

$$\text{variance} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{4} (0 + 16 + 0 + 16) = \frac{32}{4} = 8$$

This called the variance.

Note : variance squared difference of each observation from the mean.

Finding the average, just as we did before, means adding all of these values and dividing by how many there are.

In our case, we had **0, 16, 0, 16** and we dividing by **4** because we have four observations. However, this is an average of squared values which we only did to get positive values in the first place.

$$\text{variance} = 8 \Rightarrow \text{standard} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{8} = 2.83$$

Standard deviation is the square root of the variance.

So, to get our standard deviation, we take the square root of this ending value.

Here, our standard deviation is 2.83. So this on average how far each point in our data set is from the mean, which is the definition of the standard deviation.

Quiz : Measures of Spread (Calculation and Units)

QUESTION 1 OF 3

If we measure the variance associated with our sales in dollars for each month for 3 years, what are the units associated with the variance?

Dollars

Years

Dollars per Year

Dollars Squared

Dollars per Month

For the following set of data provide the value of the **variance**. (Round your answer to two decimal places at the end of your calculation. Do not round values before the final answer)

$$1, 5, 10, 3, 8, 12, 4$$

Solution :

$$\text{Variance} = \frac{(x_i - \bar{x})^2}{n}, \quad n = 7 \quad \Rightarrow \quad \bar{x} = \frac{1+5+10+3+8+12+4}{7} = 6.14$$

$$\text{Variance} = \frac{(1-6.14)^2 + (5-6.14)^2 + (10-6.14)^2 + (3-6.14)^2 + (8-6.14)^2 + (12-6.14)^2 + (4-6.14)^2}{7} = 13.55$$

For the following set of data provide the value of the **standard deviation**. (Round your answer to two decimal places at the end of your calculation.)

$$1, 5, 10, 3, 8, 12, 4$$

Solution :

$$\text{Standard deviation} = \sqrt{13.55} = 3.68$$

Introduction to the Standard Deviation and Variance

Other Measures of Spread

5 Number Summary

In the previous sections, we have seen how to calculate the values associated with the **five number summary (min, Q₁, Q₂, Q₃, max)**, as well as the measures of spread associated with these values (**range** and **IQR**).

For datasets that are **not symmetric**, the five number summary and a corresponding box-plot are a great way to get started with understanding the spread of your data. **Although I still prefer a histogram in most cases, box-plots can be easier to compare two or more groups.**

Variance and Standard Deviation

Two additional **measures of spread** that are used all the time are the **variance** and **standard deviation**. At first glance, the variance and standard deviation can seem overwhelming. If you do not understand the equations below. In this section, I just want to give you an overview of what the next sections will cover. We will walk through each of these parts thoroughly in the next few sections, **but the big picture is to gain the following:**

1. An understanding of how the mean and variance are calculated.
2. Intuition for why the calculations of the mean and variance make sense to capture the spread of our data.
3. Fields where you might see these values used.
4. Why we might use the standard deviation or variance as opposed to the values associated with the 5 number summary for a particular dataset.

Calculation

We calculate the variance in the following way:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

The variance is **the average squared difference of each observation from the mean**. The standard deviation is the square root of the variance. Therefore, the calculation of the standard deviation looks like the following:

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

The standard deviation is a measurement that has the same units as the rest of our data, and the units of the variance are the square of our original data.

Again, **this section is aimed as a precursor to the information you will be covering on the next many sections.**

Why the Standard Deviation?

It may seem odd to make this calculation of the standard deviation, that is, a complex way to measure the prevalence of our data compared to the summary of the number 5 we have seen earlier. However, it is clear that standard deviation is always used to obtain a single number compared to the spread of two sets of data.

It is good to be able to talk about how data is spread from one group to another without having to report the entire value table. We can only compare the standard deviation of one group with the standard deviation of another group.

The presence of a single number simplifies the amount of information that the person you have reported needs to consume. Having one value also has another advantage in terms of what is known as deductive statistics, but this goes beyond what we need to know now.

At the moment, we just need to know that we have a way to take all our values and get a number that tells us how to spread from one group to another.

Important Final Points

Important Final Points

1. The variance is used to compare the spread of two different groups. A set of data with higher variance is more spread out than a dataset with lower variance. Be careful though, there might just be an outlier (or outliers) that is increasing the variance, when most of the data are actually very close.
2. When comparing the spread between two datasets, the units of each must be the same.
3. When data are related to money or the economy, higher variance (or standard deviation) is associated with higher risk.
4. The standard deviation is used more often in practice than the variance, because it shares the units of the original dataset.

Use in the World

The standard deviation is associated with risk in finance, assists in determining the significance of drugs in medical studies, and measures the error of our results for predicting anything from the amount of rainfall we can expect tomorrow to your predicted commute time tomorrow.

These applications are beyond the scope of this lesson as they pertain to specific fields, but know that understanding the spread of a particular set of data is extremely important to many areas. In this lesson you mastered the calculation of the most common measures of spread.

Quiz : Advanced: Standard Deviation and Variance

QUESTION 1 OF 3

Assume d_1 and d_2 are datasets both measured in the same units. We know that the standard deviation of d_1 is 5 and the variance of d_2 is 36, which of the following are certainly true. Mark all that apply.

- The mean is larger for d_1 than for d_2 .

- The variance for d_2 is larger than for d_1 .

- The standard deviation for d_2 is larger than for d_1 .

- The median for d_2 is larger than for d_1 .

- The range for d_2 is larger than for d_1 .

QUESTION 2 OF 3

If a dataset has a standard deviation of zero, which of the following MUST be true?

- All the data points must be zero.
 - All the data points must be the same.
-

- We made a calculation error because it is not possible for the standard deviation to be zero.

QUESTION 3 OF 3

For each of the below: If the statement is **true**, mark the box next to the statement.

- If two datasets have the same variance, they will also have the same standard deviation.
-

- If I have two investment options with the same mean return, it really doesn't matter which I invest in.
-

- If I have two investment options with the same standard deviation associated with the return, they will also have the same max possible return.

Quiz : Applied Standard Deviation and Variance

Investment Data

Consider we have two investment opportunities:

	Returns					
	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Investment 1	5%	5%	5%	5%	5%	5%
Investment 2	12%	-2%	10%	0%	7%	3%

The returns for 6 consecutive years for each investment are shown above. Use this information to answer the questions below.

What is the mean return for Investment 1? (Write your answer as a percentage without the percentage sign - Ex: 9% should be reported as 9 or 93% is reported as 93.)

Solution:

$$\text{Return} = \frac{5+5+5+5+5+5}{6} = 5$$

What is the mean return for Investment 2? (Write your answer as a percentage without the percentage sign - Ex: 9% should be reported as 9 or 93% is reported as 93.)

Solution:

$$\text{Return} = \frac{12 + (-2) + 10 + 0 + 7 + 3}{6} = 5$$

Investment Data

In the previous two questions, you should have found that these investments have the same mean! That is, regardless of which investment opportunity you choose, you are expected to earn the same amount. So how are they different? Let's look at some additional questions to see if we can find some differences.

The same data as above is provided again (to minimize scrolling).

	Returns					
	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Investment 1	5%	5%	5%	5%	5%	5%
Investment 2	12%	-2%	10%	0%	7%	3%

The returns for 6 consecutive years for each investment are shown above. Use this information to answer the questions below.

QUESTION 3 OF 4

Using the information above, mark all of the below that are **true** statements.

- The risk associated with investment 1 is lower than the risk associated with Investment 2.
-
- The standard deviation associated with Investment 1 is smaller than the standard deviation associated with Investment 2.
-
- Knowing the mean return amount across all the years for each investment provides us with all of the information necessary to understand which investment we should choose.

QUESTION 4 OF 4

Based on the observed data, which of the above two investments has the best opportunity of earning more than 7%?

- Investment 1
-
- Investment 2
-
- Neither.
-
- We cannot tell.

Useful Insight

The above example is a simplified version of the real world, but does point out something useful that you may have heard before. Notice if you were not fully invested in either **Investment 1** or fully invested in **Investment 2**, but instead you were diversified across both investment options, you could earn more than either investment individually. This is the benefit of diversifying your portfolio for long term gains. For short term gains, you might not need or want to diversify. You could get lucky and hit short term gains associated with the upswings (12%, 10%, or 7%) of Investment 2. However, you might also get unlucky, and hit a down term and earn nothing or even lose money on your investment using this same strategy.

Final Quiz on Measures Spread

QUESTION 1 OF 2

For the following dataset, match each value to the appropriate label:

15, 4, 3, 8, 15, 22, 7, 9, 2, 3, 3, 12, 6

TERM	VALUE
n	13
median	7
first quartile	3
third quartile	13.5
mean	8.4
mode	3

QUESTION 2 OF 2

For the following dataset, match each value to the appropriate label:

15, 4, 3, 8, 15, 22, 7, 9, 2, 3, 3, 12, 6

TERM	VALUE
Inter quartile range	10.5
Range	20
Variance	33.9
standard deviation	5.8
minimum	2
maximum	22

Measures of Center and Spread Summary

Recap

Variable Types

We have covered a lot up to this point! We started with identifying data types as either **categorical** or **quantitative**. We then learned, we could identify quantitative variables as either **continuous** or **discrete**. We also found we could identify categorical variables as either **ordinal** or **nominal**.

Categorical Variables

When analyzing categorical variables, we commonly just look at the count or percent of a group that falls into each **level** of a category. For example, if we had two **levels** of a dog category: **lab** and **not lab**. We might say, 32% of the dogs were **lab** (percent), or we might say 32 of the 100 dogs I saw were labs (count).

However, the 4 aspects associated with describing quantitative variables are not used to describe categorical variables.

Quantitative Variables

Then we learned there are four main aspects used to describe quantitative variables:

1. Measures of **Center**
2. Measures of **Spread**
3. **Shape** of the Distribution
4. **Outliers**

We looked at calculating measures of **Center**

1. **Means**
2. **Medians**
3. **Modes**

We also looked at calculating measures of **Spread**

1. **Range**
 2. **Interquartile Range**
 3. **Standard Deviation**
 4. **Variance**
-

Calculating Variance

We saw that we could calculate the **variance** as:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

You will also see:

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Standard Deviation vs. Variance

The standard deviation is the square root of the variance. In practice, you usually use the standard deviation rather than the variance. The reason for this is because the standard deviation shares the same units with our original data, while the variance has squared units.

What Next?

In the next sections, we will be looking at the last two aspects of quantitative variables: **shape** and **outliers**. What we know about measures of center and measures of spread will assist in your understanding of these final two aspects.

Shape

Histograms

We learned how to build a **histogram**, as this is the most popular visual for quantitative data.

Shape

From a histogram we can quickly identify the shape of our data, which helps influence all of the measures we learned in the previous concepts. We learned that the distribution of our data is frequently associated with one of the three **shapes**:

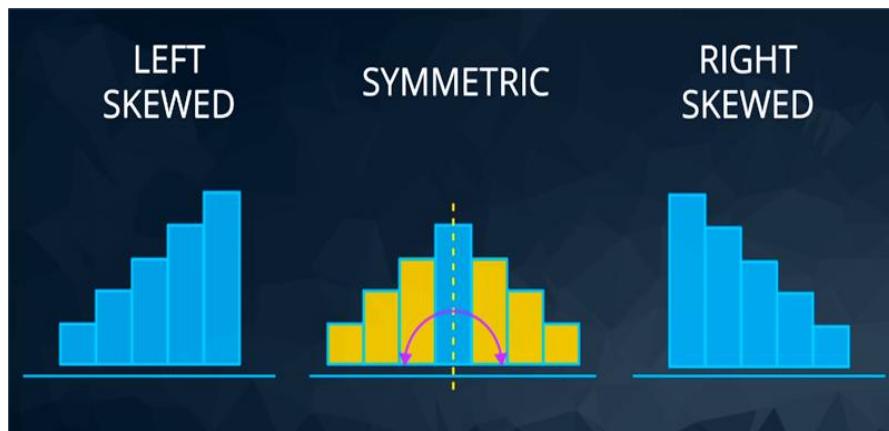
Shape	Mean vs. Median	Real World Applications
Symmetric (Normal)	Mean equals Median	Height, Weight, Errors, Precipitation
Right-skewed	Mean greater than Median	Amount of drug remaining in a blood stream, Time between phone calls at a call center, Time until light bulb dies
Left-skewed	Mean less than Median	Grades as a percentage in many universities, Age of death, Asset price changes

The mode of a distribution is essentially the tallest bar in a histogram. There may be multiple modes depending on the number of peaks in our histogram.

Here we have 3 histograms, showing the three data sets.

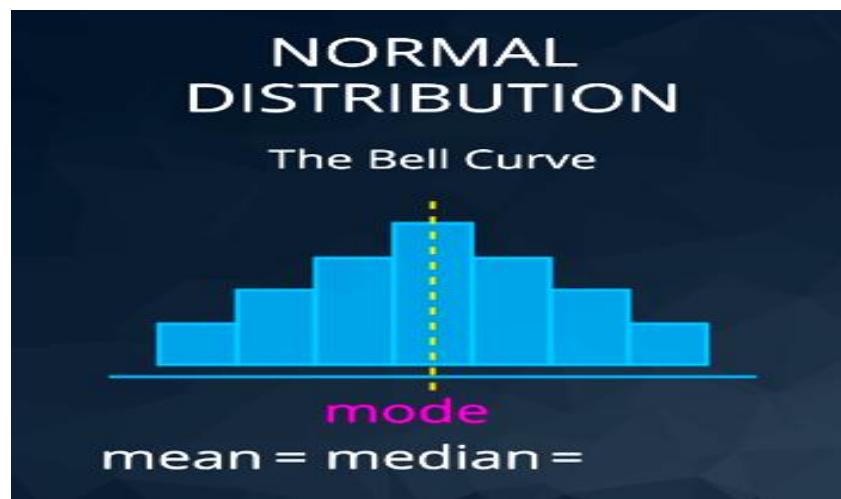


A histogram with shorter categories on the left and longer categories on the right is an oblique shape to the left. A histogram with shorter categories on the right and longer categories on the left is a distorted shape to the right. Any distribution in which we can draw a line in the middle and have the right side corresponding to the left side is symmetrical.



One of the most common symmetric distributions is known as a normal distribution, also called a hypothetical curve. The distribution pattern can tell us a lot about center measurements and diffusion.

In symmetrical distributions, such as mean is equal to the median, and is also equal to the "mode" and each of these measures exists in the center.

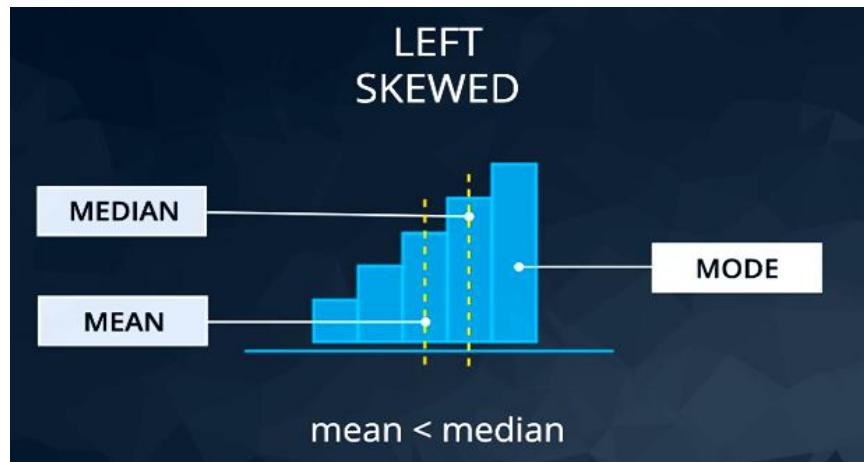


The mode is essentially the tallest bar in our histogram, when we have skewed distributions, it is the case of the mean is pulled by the tail of the distribution, while the median stays closer to the mode.

For example, in this right skewed distribution, the mean will be pulled higher, resulting in a mean that's greater than our median.

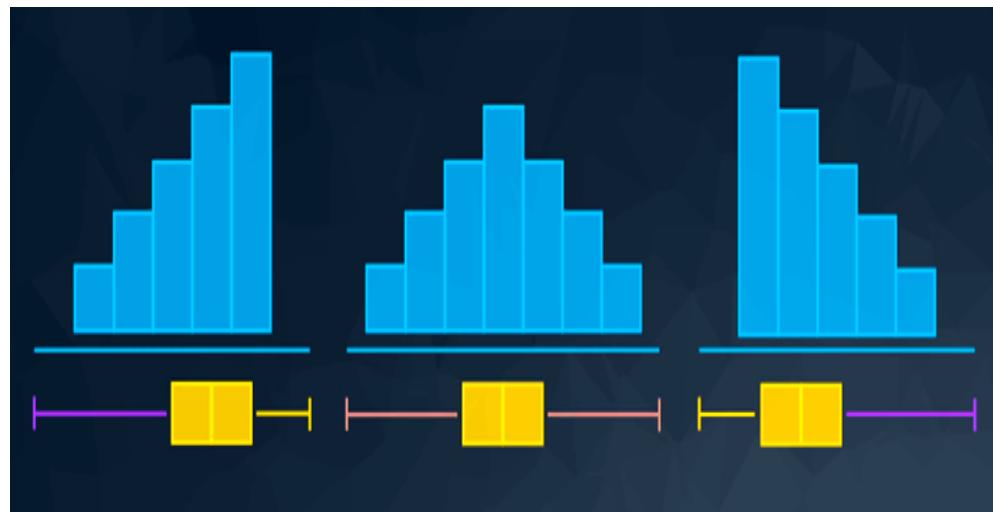


alternatively, in a left skewed distribution, our mean is pulled down here, resulting in a mean that's less than our median.



In order to relate this to the visual of a histogram, back to the 5 number summary we saw on the earlier lessons, here the corresponding box plots below each histogram.

Notice how the whiskers stretch in the direction of the skew, for each the skewed distributions. That is the longer whisker is on the left, for a left skewed distribution and it's on the right, for the right skewed distribution. Alternatively, the symmetric histogram also has a symmetric box spot.



The Shape For Data In The World

If we work with data, we can always build a quick see the shape.

Just to apply some context, some examples of bell-shaped data include approximations, weights, test scores, precipitation amounts, the mean of a distribution, or errors in manufacturing processes.

Common data that follow left skewed distributions include GPAs, age of death, and access price changes.

Common data that follow approximately right skewed distributions include the amount of drug left in your bloodstream over time, the distribution of wealth and human athletic abilities.

Though these three, Right skewed, Left skewed and symmetric, are the most common distributions, Data in the real world can be unordered and may not follow any of these distributions.

We'll talk about it more in the next section.

Quiz: Shape and Outliers (What's the Impact?)

QUESTION 1 OF 2

Match the distribution shape with the correct relationship in comparing the mean to the median.

Cannot be Determined.

SHAPE	COMPARISON
Right-skewed	Mean is greater than the Median.
Left-skewed	Mean is less than the Median.
Symmetric	Mean is equal to the Median.

QUESTION 2 OF 2

Check all of the below that must be true.

- For a dataset the mean equals the median, so the data are normally distributed.
- Normally distributed data must have a mean equal the median.
- All data are normally distributed.
- Data must be distributed either right-skewed, left-skewed, or symmetric.
- Histograms and box plots are both used to plot quantitative data. They cannot be used to plot categorical data.
- A box plot relates directly to the 5 number summary.

Shape and Outliers

Outliers

We learned that **outliers** are points that fall very far from the rest of our data points. This influences measures like the mean and standard deviation much more than measures associated with the five number summary.

Identifying Outliers

There are a number of different techniques for identifying outliers.

In this lesson we want to take a look at the last aspect used to describe quantitative variables and outliers .

Outliers : data points that fall very far from the rest of the values in our dataset.

To determine the points too far, there are a number of different ways.

Usually we look at a histogram and see if the point is too far from other data points. Again, a quick plot of your data can often helps us to understand a lot in a short amount of time.

In order to illustrate the impact that outliers can have on the way we report summary statistics, let's consider the salaries of entrepreneurs.

Imagine we select 10 entrepreneurs earning and we pull these 9 values here as earnings in thousands of dollars (**45, 68, 92, 53, 105, 56, 24, 15, 155**), and the tenth is the CEO of Facebook (**\$1.6 Billion**). According to a post by CNN in 2016, he earned **4.4 million** dollars per day.

Here, we can calculate the mean of these salaries for entrepreneurs based in this data to be approximately **160 million** dollars. This is incredibly misleading.

Literally zero of the entrepreneurs earned this salary. None of the 10 salaries are even close to this amount. A better measure of center would certainly be the median.

The median here is calculated at 62000 dollars a year and is a batter indication of what an entrepreneur is likely to earn based in our data.

Our standard deviation is also not a great measure in this case. At approximately 482 million dollars, all this suggests is that our earning for entrepreneurs are really spread out, but that really isn't fair either. Just one point is really far from the rest. Like really really far.



Working With Outliers

Common Techniques

When outliers are present we should consider the following points.

1. Noting they exist and the impact on summary statistics.

In this case, our mean and standard deviation greatly increases. If the outliers are typos or data entry errors, this is reason to remove these points.

2. If typo - remove or fix.

If we know what it should be, we can change it using the correct values. In similar cases in the example above, we may try to understand what was quite different about the outlier when compared to other individuals.

3. Understanding why they exist, and the impact on questions we are trying to answer about our data.

There is a whole field aimed at this idea delineating the discovery of abnormal values.

We will note that the measurements associated with the sum of 5, such as the minimum, the maximum, and the first, second, and third quadrants are more useful when the abnormal values exist.

4. Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.

This example shows us that we need to be careful about how our results are shared and to report our conclusions using summary statistics when we have outlier values.

5. Be careful in reporting. Know how to ask the right questions.

Some statistics are more misleading than others. If we are consumers of data based on information, on which we all rely, it is important to know how to ask questions about statistics around us.

Working With Outliers My Advice

Outliers Advice

Below are my guidelines for working with any column (random variable) in your dataset.

1. Plot your data to identify if you have outliers.
 2. Handle outliers accordingly via the methods above.
 3. If no outliers and your data follow a normal distribution - use the mean and standard deviation to describe your dataset, and report that the data are normally distributed.
-

Side note

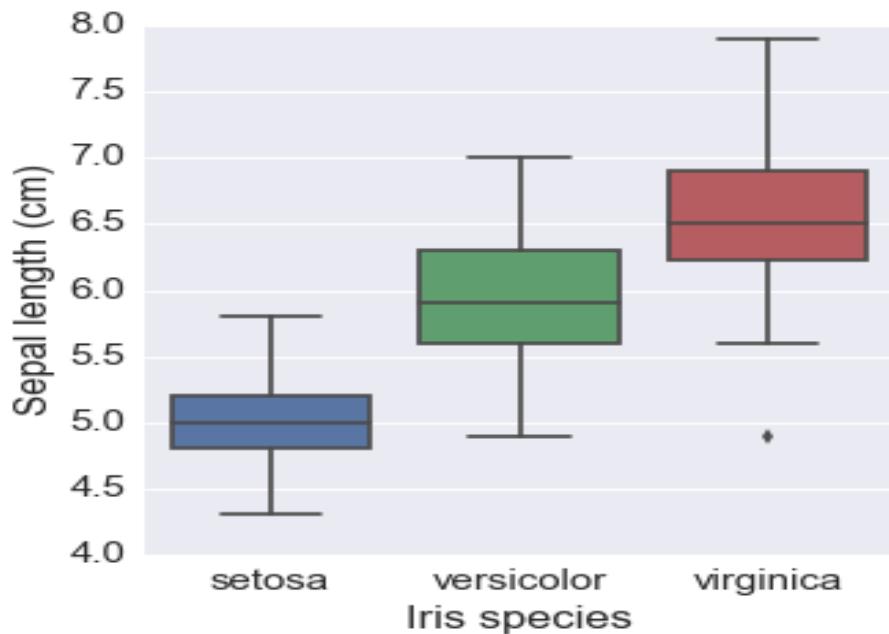
If you aren't sure if your data are normally distributed, there are plots called [normal quantile plots](#) and statistical methods like. Implementing this test is beyond the scope of this class, but can be used as a fun fact.

4. If you have skewed data or outliers, use the five number summary to summarize your data and report the outliers.

Quiz: Shape and Outliers (Comparing Distributions)

Image Summary

In the below image, we have three box-plots. Each box-plot is for a different Iris flower: **setosa**, **versicolor**, or **virginica**. On the y-axis, we are given the sepal length. Notice that **virginica** has an **outlier** towards the bottom of the plot. Therefore, the **minimum** is not given by the bottom line here; rather, it is provided by this point.



	SETOSA	VERSICOLOR	VIRGINICA
RANGE = MAX - MIN	$5.8 - 4.3 = 1.5$	$7 - 4.9 = 2.1$	$7.9 - 5.6 = 2.3$
INTERQUARTIEL RANGE $IQR = Q_3 - Q_1$	$5.2 - 4.8 = 0.4$	$6.3 - 5.6 = 0.7$	$6.9 - 6.3 = 0.6$
MEDIAN " Q_2 "	5	5.8	6.5
THIRDQUARTILE " Q_3 "	5.2	6.3	6.5
APPROXIMATELY SYMMETRIC			
LARGEST SEPAL	5.8	7	7.8

QUESTION 1 OF 2

Match the appropriate Iris type to the statement(s) that are true for its Sepal Length.

None

SEPAL LENGTH	IRIS TYPE
The largest Range	Virginica
The smallest Interquartile Range	Setosa
Median is approximately 5	Setosa
Third quartile is approximately 6.3	Versicolor
Approximately Symmetric	All
The largest sepals on average.	Virginica

QUESTION 2 OF 2

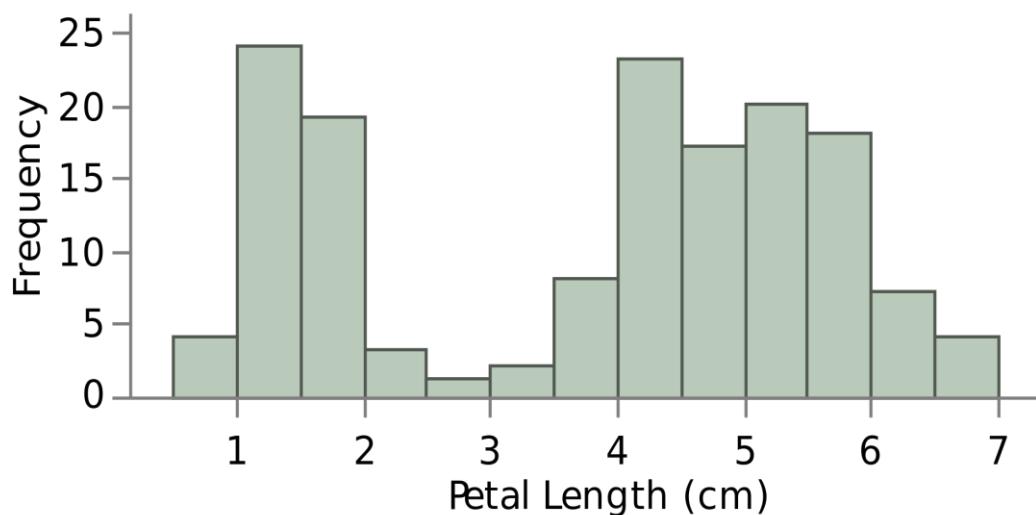
Using the same flower data, select all of the below statements that **MUST** be true.

- All setosa flowers have a shorter sepal length than versicolor flowers.
- All virginica flowers have larger sepal length than setosa flowers.
- More than 75% of the virginica flowers have a larger sepal length than the largest setosa flower.
- More than 50% of setosa flowers have larger sepal length than the shortest versicolor flower.
- More data was collected on versicolor flowers than on virginica flowers.
- More data was collected on setosa flowers than on virginica flowers.

Quiz: Shape and Outliers (Visuals)

Questions 1 - 2: Petal Length

The below plot will be used to answer the first two questions in this section.



QUESTION 1 OF 5

What is the name of the above plot?

Bar Chart

Box Plot

Histogram

Pie Chart

QUESTION 2 OF 5

What is the shape of the above distribution?

Right skewed

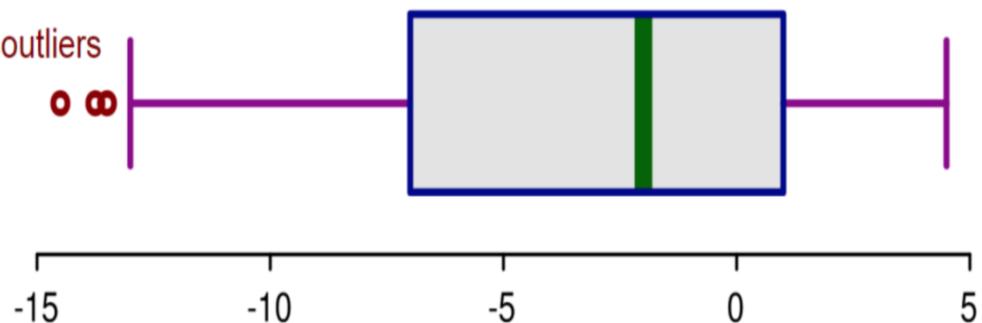
Left skewed

Symmetric

Bi-modal

Questions 3 - 6: Shape and Outliers

Use the below image to assist with answering the next three questions.



QUESTION 3 OF 5

What is the name of the above plot?

Bar Chart

Box Plot

Histogram

Pie Chart

QUESTION 4 OF 5

What is the shape of the distribution?

Right skewed

Left skewed

Symmetric

Bi-modal

QUESTION 5 OF 5

Select the true statement for the box-plot above.

The mean is less than the median.

The mean is greater than the median.

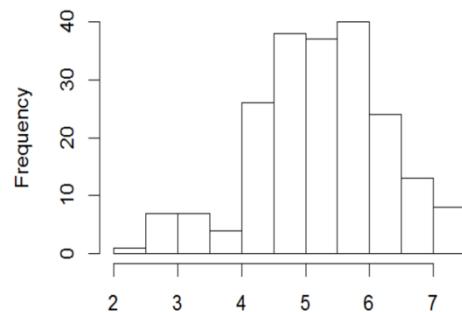
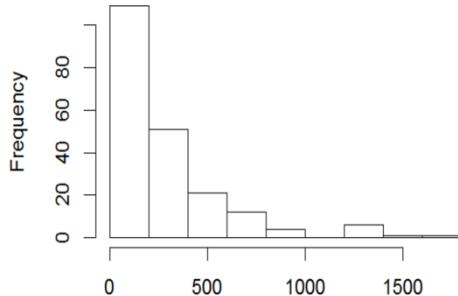
The mean is approximately equal to the median.

It is impossible to tell the relationship between the mean and median.

Quiz: Shape and Outliers (Final Quiz)

Histograms

Let the histogram on the left be **Histogram 1** and the histogram on the right be **Histogram 2**.



QUIZ QUESTION

Correctly match the histograms to the statements that are true about each.

Histogram 2

Neither

Impossible to tell.

Histogram 1

Both

STATEMENT

HISTOGRAM

Mean is greater than the median.

Histogram 1

Data has higher variance.

Histogram 1

Binwidth is equal to 0.5.

Histogram 2

The range is approximately 5.5.

Histogram 2

Distribution is left-skewed.

Neither

The mean is approximately equal to the median.

Histogram 2

Descriptive Statistics Summary

Recap

Variable Types

We have covered a lot up to this point! We started with identifying data types as either **categorical** or **quantitative**. We then learned, we could identify quantitative variables as either **continuous** or **discrete**. We also found we could identify categorical variables as either **ordinal** or **nominal**.

Categorical Variables

When analyzing categorical variables, we commonly just look at the count or percent of a group that falls into each **level** of a category. For example, if we had two **levels** of a dog category: **lab** and **not lab**. We might say, 32% of the dogs were **lab** (percent), or we might say 32 of the 100 dogs I saw were labs (count).

However, the 4 aspects associated with describing quantitative variables are not used to describe categorical variables.

Quantitative Variables

Then we learned there are four main aspects used to describe **quantitative** variables:

1. Measures of **Center**
 2. Measures of **Spread**
 3. **Shape** of the Distribution
 4. **Outliers**
-

Measures of Center

We looked at calculating measures of **Center**

1. Means
 2. Medians
 3. Modes
-

Measures of Spread

We also looked at calculating measures of **Spread**

1. Range
 2. Interquartile Range
 3. Standard Deviation
 4. Variance
-

Shape

We learned that the distribution of our data is frequently associated with one of the three **shapes**:

1. Right-skewed
2. Left-skewed
3. Symmetric (frequently normally distributed)

Depending on the shape associated with our dataset, certain measures of center or spread may be better for summarizing our dataset.

When we have data that follows a **normal** distribution, we can completely understand our dataset using the **mean** and **standard deviation**.

However, if our dataset is **skewed**, the **5 number summary** (and measures of center associated with it) might be better to summarize our dataset.

Outliers

We learned that outliers have a larger influence on measures like the mean than on measures like the median. We learned that we should work with outliers on a situation by situation basis. Common techniques include:

1. At least note they exist and the impact on summary statistics.
 2. If typo - remove or fix
 3. Understand why they exist, and the impact on questions we are trying to answer about our data.
 4. Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.
 5. Be careful in reporting. Know how to ask the right questions.
-

Histograms and Box Plots

We also looked at histograms and box plots to visualize our quantitative data. Identifying outliers and the shape associated with the distribution of our data are easier when using a visual as opposed to using summary statistics.

What Next?

Up to this point, we have only looked at **Descriptive Statistics**, because we are describing our collected data. In the final sections of this lesson, we will be looking at the difference between **Descriptive Statistics and Inferential Statistics**.

Descriptive vs. Inferential Statistics

In this section, we learned about how **Inferential Statistics** differs from **Descriptive Statistics**.

Descriptive statistics is about describing our collected data.

Inferential Statistics

Inferential Statistics is about using our collected data to draw conclusions to a larger population.

We looked at specific examples that allowed us to identify the

- 1. Population** - our entire group of interest.
- 2. Parameter** - numeric summary about a population
- 3. Sample** - subset of the population
- 4. Statistic** numeric summary about a sample

Example :

Imagine that we want to deduce the proportion of all students of the udasity Academy who drink coffee.

We send an e-mail to all of udasity's graduates and current students asking the question: **Do you drink coffee?**

Let's say that a list contained 100,000 emails.

Unfortunately, not everyone we e-mailed responded, and then we received only 5,000 replies.

We found that 73% of people who responded to the e-mail that we sent to everyone said they did not drink coffee.

Note : Metadata is about describing our data. This means that any information we have and share with regard to the 5000 responses is descriptive.

In our example, the conclusion statistics are to draw conclusions about the coffee habit followed by all of the 100,000 Udasite students using 5,000 responses from our sample.

The general language associated with this scenario is as described here.

POPULATION	= 100,000 students
SAMPLE	= 5000 students
STATISTIC	= 73%
PARAMETER	= Proportion of all 100,000 students that drink coffee

We have a group that represents our entire group. In our case, we mean 100,000 students.

We collect a subset of this group, which we call the sample name. In the case we have, we mean 5,000 students.

Any numerical summary calculated from the sample is called the census name. In our case, we mean 73% of the 5,000 students who drink coffee. This figure represents 73%.

The group's Digest is known as the parameter. In our case, we are not familiar with this value because it is a number that requires information from all students of Udacity.

The conclusions about the parameter based on the statistics are known as the conclusion.

Quiz: Descriptive vs. Inferential (Udacity Students)

QUIZ QUESTION

Identify the population, parameter, sample, and statistic for the below scenario:

Consider we are interested in the average number of hours slept by all Udacity students (100,000 students). I send an email to all Udacity students, but I only receive 5,000 response emails. The average amount of sleep of those that responded was 6.8 hours of sleep.

6.8 hours of sleep	We cannot know for sure.
--------------------	--------------------------

TERM	DESCRIPTION
Population	All Udacity students
Parameter	We cannot know for sure.
Sample	5,000 Udacity students
Statistic	6.8 hours of sleep

Quiz: Descriptive vs. Inferential (Bagels)

QUESTION 1 OF 2

Identify the population(s), parameter(s), sample(s), and statistic(s) for the below scenario:

Consider we own a bagel shop. We know that the average diameter of all of our bagels is 5.5 inches. A competitor moves right next door to us! We are interested in if they make larger bagels than us. We obtain 100 of their bagels, and we find they have an average diameter of 6 inches.

Statistic **Parameter** **None** **Sample**

DESCRIPTION	TERM
5.5 inches	Parameter
6 inches	Statistic
All the bagels at our bagel shop.	Population
All the bagels at our competitor's bagel shop.	Population
The 100 bagels from the competitor's bagel shop.	Sample

QUESTION 2 OF 2

For the below, match the term to the correct description.

Statistic None Statistic Population

DESCRIPTION	TERM
A numeric summary of a sample.	Statistic
A numeric summary of a population.	Parameter
Drawing conclusions regarding a population using information from a sample.	Inference
Drawing conclusions regarding a sample using information from a population.	None
A subset of a population.	Sample
Our entire group of interest.	Population
Frequently we do not know this value, so we must try and estimate.	Parameter

Descriptive vs. Inferential Summary

Descriptive vs. Inferential Statistics

In this section, we learned about how **Inferential Statistics** differs from **Descriptive Statistics**.

Descriptive Statistics

Descriptive statistics is about describing our collected data using the measures discussed throughout this lesson: measures of center, measures of spread, shape of our distribution, and outliers. We can also use plots of our data to gain a better understanding.

Inferential Statistics

Inferential Statistics is about using our collected data to draw conclusions to a larger population. Performing inferential statistics well requires that we take a sample that accurately represents our population of interest.

A common way to collect data is via a survey. However, surveys may be extremely biased depending on the types of questions that are asked, and the way the questions are asked. This is a topic you should think about when tackling the first project.

We looked at specific examples that allowed us to identify the

1. **Population** - our entire group of interest.
 2. **Parameter** - numeric summary about a population
 3. **Sample** - subset of the population
 4. **Statistic** numeric summary about a sample
-

Looking Ahead

Though we will not be diving deep into inferential statistics within this course, you are now aware of the difference between these two avenues of statistics. If you have ever conducted a hypothesis test or built a confidence interval, you have performed inferential statistics. The way we perform inferential statistics is changing as technology evolves. Many career paths involving **Machine Learning** and **Artificial Intelligence** are aimed at using collected data to draw conclusions about entire populations at an individual level. It is an exciting time to be a part of this space, and you are now well on your way to joining the other practitioners!

SQL Introduction

There are many uses for SQL in software development. We will focus on the parts of SQL used in data analysis.

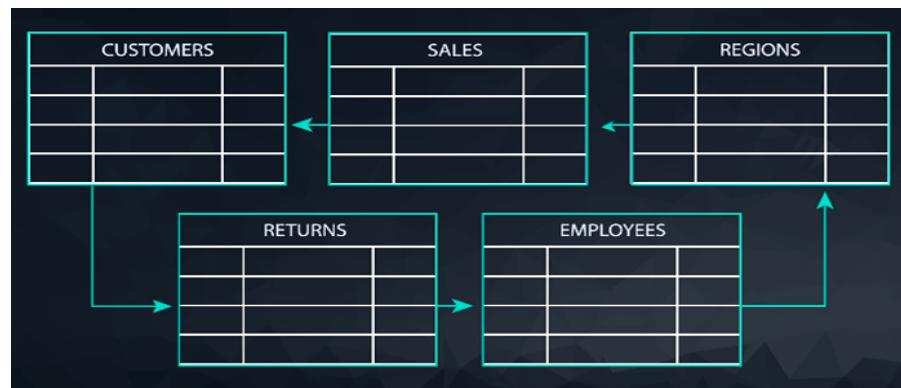
At the moment, companies are gathering a lot of data and expect staff in all fields to use it systematically in decision making. There are many ways to do this.

The Parch & Posey Database

For examples of this course, we will use data from Parch & Posey, a paper-selling company. It has 50 sales representatives spread across the United States in 4 regions. It sells 3 types of paper, plain, poster and glossy. And their clients are the largest of the 100 richest companies who attract them by advertising on Google, Facebook and Twitter.

By using SQL, we will be able to help Parch & Posey answer the deceptive questions such as: Which of their product lines is worst performing ? Which of their they should make a greater marketing investment in?

There is one way to store data and is on a spreadsheet. Sometimes, we need a lot of spreadsheets to store data from different sources. We can represent relationships between spreadsheets by using Entity relationship diagram "**ERD**".



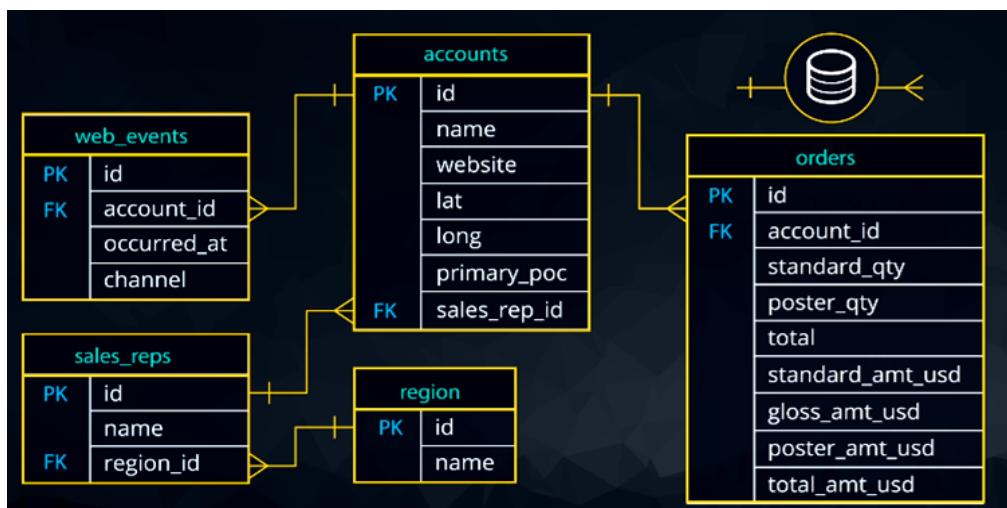
Entity Relationship Diagrams

An **entity relationship diagram** (ERD) is a common way to view data in a database.

Below is the ERD for the database we will use from Parch & Posey. **These diagrams help you visualize the data you are analyzing including:**

1. The names of the tables.
2. The columns in each table.
3. The way the tables work together.

This is what an ERD looks like.



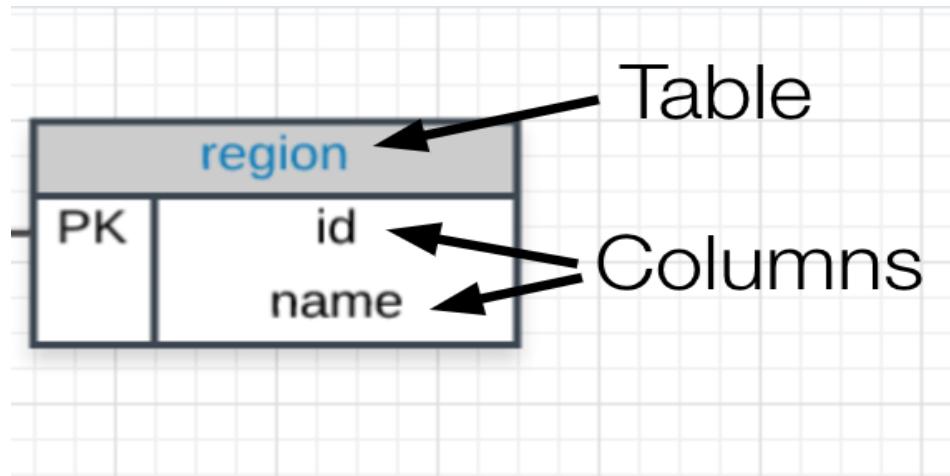
Each spreadsheet is represented on a table. At the top, you will see the name of the table, and then below, each column name is listed.

What to Notice

In the Parch & Posey database there are five tables (essentially 5 spreadsheets):

1. **web_events**
2. **accounts**
3. **orders**
4. **sales_reps**
5. **region**

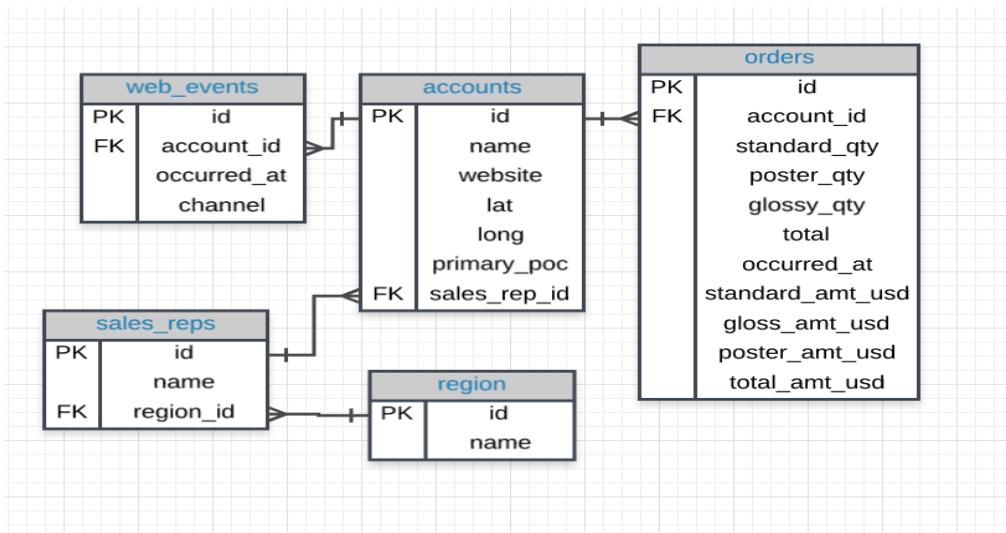
You can think of each of these tables as an individual spreadsheet. Then the columns in each spreadsheet are listed below the table name. For example, the **region** table has two columns: **id** and **name**. Alternatively the **web_events** table has four columns.



The "crow's foot" that connects the tables together shows us how the columns in one table relate to the columns in another table. In this first lesson, you will be learning the basics of how to work with SQL to interact with a single table. In the next lesson, you will learn more about why these connections are so important for working with SQL and relational databases.

- SQL is a language used to interact with a database. It can query one table, or it can query across multiple tables.
- Understanding these relationships will help you gain insights from your data faster than with other software.

Quiz: ERD Fundamentals



QUIZ QUESTION

Use the above image and your new knowledge of ERDs to match each **Term** to the appropriate **Definition**.

Submit to check your answer choices!

DEFINITION

TERM

A column name in the Parch & Posey database.

primary_poc

A table name in the Parch & Posey database.

web_events

A collection of tables that share connected data

Database

stored in a computer.

A diagram that shows how data is structured in

ERD

a database.

A language that allows us to access data stored

SQL

in a database.

Map of SQL Content

Introduction

Throughout the next three lessons, you will be learning how to write **Structured Query Language (SQL)** to interact with a database here in the classroom. You will not need to download any software, and you will still be able to test your skills!

SQL is an extremely in demand skill. [Tons of jobs use SQL](#), and in the next lessons you will be learning how to utilize SQL to analyze data and answer business questions.

Project

The skills you learn in the classroom are directly extendable to writing **SQL** in other environments outside this classroom. For the project at the end of these lessons, you will download a program that will allow you to write code on your local machine. You will then analyze and answer business questions using data associated with a music store by querying their database.

Lesson Outline

There are three lessons in this Nanodegree aimed at helping you understand how to write SQL queries. If you choose to take the [Business Analyst Nanodegree](#) or the [Data Analyst Nanodegree](#) programs, these three lessons will also be a part of these programs. However, there is also an additional lesson on Advanced SQL also taught by Derek!

The three lessons in this course aim at the following components of SQL:

- **SQL Basics** - Here you will get your first taste at how SQL works, and learn the basics of the SQL language. You will learn how to write code to interact with tables similar to the ones we analyzed in Excel earlier. Specifically, you will learn a little about databases, the basic syntax of SQL, and you will write your first queries!

- **SQL Joins** - In this lesson, you will learn the real power of SQL. You will learn about Entity Relationship Diagrams (ERDs), and how to join multiple tables together from a relational database. The power to join tables is what really moved companies to adopt this approach to holding data.
- **SQL Aggregations** - In this final lesson, you will learn some more advanced features of SQL. You will gain the ability to summarize data from multiple tables in a database.

At the end of these three lessons, you will be ready to tackle the project. The project aims to assure you have mastered these three topics, but you will also see some of the more advanced queries that were not covered in this course. These are just meant to introduce you to the advanced material, but don't feel discouraged if you didn't get these - they were beyond the scope of the class, and they are not required to pass the project!

Why SQL

Introduction

Before we dive into writing SQL queries, let's take a look at what makes SQL and the databases that utilize SQL so popular.

I think it is an important distinction to say that SQL is a **language**. Hence, the last word of **SQL** being **language**. SQL is used all over the place beyond the databases we will utilize in this class. With that being said, SQL is most popular for its interaction with databases. For this class, you can think of a **database** as a bunch of excel spreadsheets all sitting in one place. Not all databases are a bunch of excel spreadsheets sitting in one place, but it is a reasonable idea for this class.

Why Do Data Analysts Use SQL?

There are some major advantages to using **traditional relational databases**, which we interact with using SQL. **The five most apparent are:**

- SQL is easy to understand.
- Traditional databases allow us to access data directly.
- Traditional databases allow us to audit and replicate our data.
- SQL is a great tool for analyzing multiple tables at once.
- SQL allows you to analyze more complex questions than dashboard tools like Google Analytics.

You will experience these advantages first hand, as we learn to write SQL to interact with data.

SQL vs. No.SQL

You may have heard of NoSQL, which stands for not only SQL. Databases using No.SQL allow for you to write code that interacts with the data a bit differently than what we will do in this course. These No.SQL environments tend to be particularly popular for web based data, but less popular for data that lives in spreadsheets the way we have been analyzing data up to this point. One of the most popular No.SQL languages is called [MongoDB](#).

Why Do Businesses Choose SQL?

1. **Data integrity is ensured** - only the data you want entered is entered, and only certain users are able to enter data into the database.
2. **Data can be accessed quickly** - SQL allows you to obtain results very quickly from the data stored in a database. Code can be optimized to quickly pull results.
3. **Data is easily shared** - multiple individuals can access data stored in a database, and the data is the same for all users allowing for consistent results for anyone with access to your database.

Databases have a number of features that make them great at accessing data. More importantly, databases help to verify the integrity of data, and can ensure consistency of data entered.

For example, if we record the number of children in a person, the database will only be able to impose integers, because we can not have half a child.

In fact, they are also databases, and they can run very quickly across very large sets of data that are ideal for additional speed.

The databases represent common entities. Many people can access a database at the same time. More importantly, all these people will deal with the same data.

How Databases Store Data

A few key points about data stored in SQL databases:

1. Data in databases is stored in tables that can be thought of just like Excel spreadsheets.

For the most part, you can think of a database as a bunch of Excel spreadsheets. Each spreadsheet has rows and columns. Where each row holds data on a transaction, a person, a company, etc., while each column holds data pertaining to a particular aspect of one of the rows you care about like a name, location, a unique id, etc.

2. All the data in the same column must match in terms of data type.

An entire column is considered quantitative, discrete, or as some sort of string. This means if you have one row with a string in a particular column, the entire column might change to a text data type. **This can be very bad if you want to do math with this column!**

3. Consistent column types are one of the main reasons working with databases is fast.

Often databases hold a LOT of data. So, knowing that the columns are all of the same type of data means that obtaining data from a database can still be fast.

Text + Quiz: Types of Databases

Types of Databases

SQL Databases

There are many different types of SQL databases designed for different purposes. In this course we will use [Postgres](#) within the classroom, which is a popular open-source database with a very complete library of analytical functions.

Some of the most popular databases include:

1. MySQL
2. Access
3. Oracle
4. Microsoft SQL Server
5. Postgres

You can also write SQL within other programming frameworks like Python, Scala, and HaDoop.

Small Differences

Each of these SQL databases may have subtle differences in syntax and available functions -- for example, MySQL doesn't have some of the functions for modifying dates as Postgres. **Most** of what you see with Postgres will be directly applicable to using SQL in other frameworks and database environments. For the differences that do exist, you should check the documentation. Most SQL environments have great documentation online that you can easily access with a quick Google search.

The article here compares three of the most common types of SQL: SQLite, PostgreSQL, and MySQL.

Though you will use PostgreSQL in the classroom, you will utilize SQLite for the project. Again, once you have learned how to write SQL in one environment, the skills are mostly transferable. So with that, let's jump in!

QUIZ QUESTION

Check all of the below that are true about your learning experience here in the classroom.

- The code you write in the classroom is exactly what you would write to analyze data in Postgre SQL.
- The code you write in the classroom is exactly what you would write to analyze data in MySQL.
- The code you write in the classroom will allow you to easily pick up any SQL programming including Microsoft SQL Server, Oracle, or SQLite.
- When you leave the classroom, you will have to re-learn how to use SQL, as it works differently outside the classroom.

Types of Statements

The key to SQL is understanding **statements**. A few statements include:

1. **CREATE TABLE** is a statement that creates a new table in a database.
2. **DROP TABLE** is a statement that removes a table in a database.
3. **SELECT** allows you to read data and display it. This is called a **query**.

The **SELECT** statement is the common statement used by analysts, and you will be learning all about them throughout this course!

Quiz: Statements

QUESTION 1 OF 2

In SQL, you can think of a statement as (select all that apply):

- A piece of correctly written SQL code.
- A way to manipulate data stored in a database.
- A sentence.
- A way to read data stored in a database.

QUESTION 2 OF 2

In the following lessons we will focus on which type of query?

- DROP
- CREATE
- QUERY
- SELECT

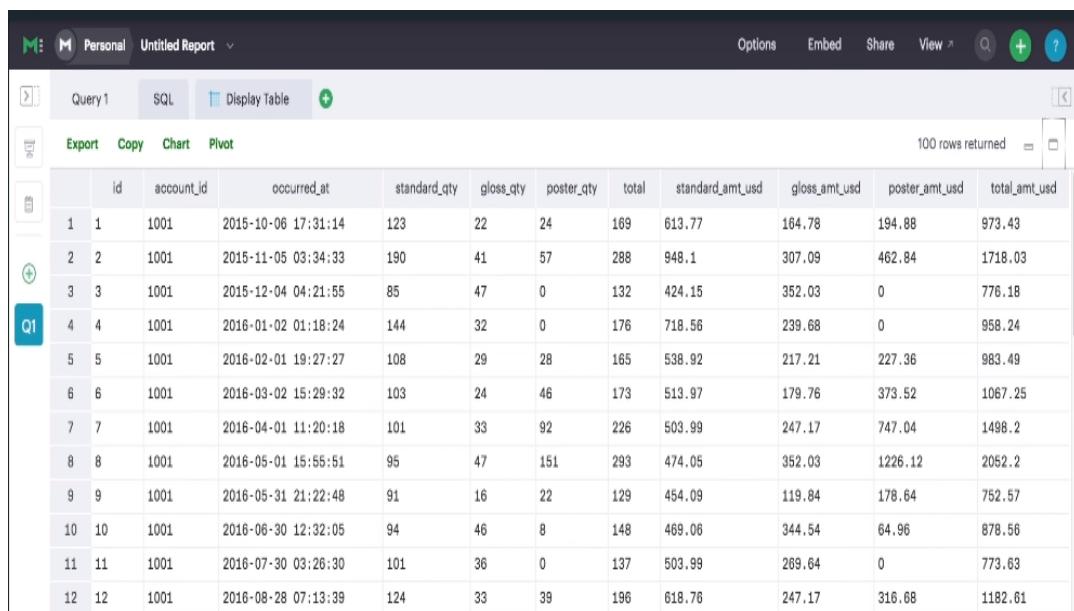
SELECT & FROM

Here you were introduced to two statements that will be used in every query you write

1. **SELECT** is where you tell the query what columns you want back.
2. **FROM** is where you tell the query what table you are querying from. Notice the columns need to exist in this table.

You will use these two statements in every query in this course, but you will be learning a few additional statements and operators that can be used along with them to ask more advanced questions of your data.

For example, **Parch & Posey** recorded their requests in a table called the Orders table. This table contains all the columns we see here , each corresponding to an attribute, or a request.



The screenshot shows a data visualization interface with a table titled "Query1". The table has 12 columns: id, account_id, occurred_at, standard_qty, gloss_qty, poster_qty, total, standard_amt_usd, gloss_amt_usd, poster_amt_usd, and total_amt_usd. There are 100 rows returned. The first few rows of data are:

1	1	1001	2015-10-06 17:31:14	123	22	24	169	613.77	184.78	194.88	973.43
2	2	1001	2015-11-05 03:34:33	190	41	57	288	948.1	307.09	462.84	1718.03
3	3	1001	2015-12-04 04:21:55	85	47	0	132	424.15	352.03	0	776.18
Q1	4	1001	2016-01-02 01:18:24	144	32	0	176	718.56	239.68	0	958.24
5	5	1001	2016-02-01 19:27:27	108	29	28	165	538.92	217.21	227.36	983.49
6	6	1001	2016-03-02 15:29:32	103	24	46	173	513.97	179.76	373.52	1067.25
7	7	1001	2016-04-01 11:20:18	101	33	92	226	503.99	247.17	747.04	1498.2
8	8	1001	2016-05-01 15:55:51	95	47	151	293	474.05	352.03	1226.12	2052.2
9	9	1001	2016-05-31 21:22:48	91	16	22	129	454.09	119.84	178.64	752.57
10	10	1001	2016-06-30 12:32:05	94	46	8	148	469.06	344.54	64.96	878.56
11	11	1001	2016-07-30 03:26:30	101	36	0	137	503.99	289.64	0	773.63
12	12	1001	2016-08-28 07:13:39	124	33	39	196	618.76	247.17	316.68	1182.61

Each request has a unique identifier, so we can refer to it and a time stamp indicating when this request was made. It also contains the quantity sold, yield and output for each type of paper requested.

To create all the requests, we write the SELECT statement.

Consider the term "Select" when filling out a form to get the results we are trying to find.

The form contains a set of questions such as:

What data do you want to pull from?

Which elements from the database do you want to pull?

These questions are structured in the same order each time. Some of which are mandatory and the other optional.

The term "query" tells you what data to use from a table. The Select statement tells you which columns to read from that table.

Column names are separated by commas, with no comma after the last column name. We can also specify all column names using Select with the asterisk "star means all columns".

The terms Select and From are mandatory. They will be included in each query we write as part of this course. If we do not guarantee both, we will get an error message instead of the desired results.

When you click the play button, the database will execute this query and the results will be returned. Each row in the order table will show all available columns.

The screenshot shows a database interface with a dark theme. At the top, there are navigation icons for 'Personal' and 'Untitled Report'. On the right, there are options for 'Options', 'Embed', 'Share', 'View', and a search icon. Below the header, there are tabs for 'Query 1', 'SQL', and 'Display Table'. The SQL tab is selected, showing the following code:

```
1 SELECT *
2 FROM demo.orders
```

Below the code, there is a preview area labeled 'Q1' containing a table with 100 rows returned. The table has the following columns: id, account_id, occurred_at, standard_qty, gloss_qty, poster_qty, total, standard_amt_usd, gloss_amt_usd, poster_amt_usd, and total_amt_usd. The data is as follows:

			occurred_at	standard_qty	gloss_qty	poster_qty	total	standard_amt_usd	gloss_amt_usd	poster_amt_usd	total_amt_usd
1	1	1001	2015-10-06 17:31:14	123	22	24	169	613.77	164.78	194.88	973.43
2	2	1001	2015-11-05 03:34:33	190	41	57	288	948.1	307.09	462.84	1718.03
3	3	1001	2015-12-04 04:21:55	85	47	0	132	424.15	352.03	0	776.18
4	4	1001	2016-01-02 01:18:24	144	32	0	176	718.56	239.68	0	958.24
5	5	1001	2016-02-01 19:27:27	108	29	28	165	538.92	217.21	227.36	983.49
6	6	1001	2016-03-02 15:29:32	103	24	46	173	513.97	179.76	373.52	1067.25
7	7	1001	2016-04-01 11:20:18	101	33	92	226	503.99	247.17	747.04	1498.2

A green success message 'Succeeded in 1s' is displayed at the bottom right of the preview area. At the bottom of the interface, there are buttons for 'Export', 'Copy', 'Chart', and 'Pivot', and a status bar indicating '100 rows returned'.

Text + Quiz: Your First Query

First SQL Statement

The tables from the Parch & Posey database are stored in the backend of the box below. You will notice on the left margin under **SCHEMA** is a list of tables that were shown in the ERD earlier. We'll write queries and run them to see the results using the table below in the same way you would in most other database environments. In this lesson, we'll access just one table at a time, and in later lessons we'll be joining and doing aggregations across tables.

The Udacity SQL Environment

In order to get started, try running the query you saw in the previous lesson ! In the environment at the bottom of the page, you have the ability to test your SQL code. In the left panel, you will find the tables that we saw earlier in the ERD. In the right panel you can write your SQL code, and we can click the **EVALUATE** button to run the query. This may take a moment to run.

The **HISTORY** menu will show your previously run queries. The **MENU** will allow you to remove the SCHEMA from the left panel, as well as reset the database.

```
SELECT *  
FROM orders;
```

You will notice that Derek was using a **demo** table (and he will continue to do this in future lessons), but you should write your queries using the table names exactly as shown in the left margin. These tables are identical to those that you see from Derek in the future lessons too (with **demo** removed).

Every query you write will have at least these two parts: SELECT and FROM. In order to evaluate your query, you can either click **EVALUATE** OR hit **control + Enter**. If you get an error, it will sometimes cover the evaluate button, so this second option is very nice! Again, **control + Enter** will run your query!

The screenshot shows a SQL query editor interface. On the left, there's a sidebar titled "Input" with a "SCHEMA" dropdown containing tables: "accounts", "orders", "region", "sales_reps", and "web_events". Below the schema is a message "Success!". On the right, the main area has "HISTORY" and "MENU" buttons at the top. The code input field contains:

```

1  SELECT *
2  FROM orders

```

Below the code, there's an "EVALUATE" button. Under the "Output" section, it says "6912 results" and includes a "Download CSV" link. The output table has columns: id, account_id, occurred_at, standard_qty, gloss_qty, and poster_qty. The data shows five rows for account_id 1001.

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1	1001	2015-10-06T17:31:14.000Z	123	22	24
2	1001	2015-11-05T03:34:33.000Z	190	41	57
3	1001	2015-12-04T04:21:55.000Z	85	47	0
4	1001	2016-01-02T01:18:24.000Z	144	32	0
5	1001	2016-02-01T19:27:27.000Z	108	29	28

Every SQL Query

Every query will have at least a **SELECT** and **FROM** statement. The **SELECT** statement is where you put the **columns** for which you would like to show the data. The **FROM** statement is where you put the **tables** from which you would like to pull data.

Try writing your own query to select only the **id**, **account_id**, and **occurred_at** columns for all orders in the **orders** table.

Solution to Previous Concept

```
SELECT id, account_id, occurred_at  
FROM orders
```

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

1 SELECT id,account_id,occurred_at
2 FROM orders

Success!

EVALUATE

Output 6912 results

id	account_id	occurred_at
1	1001	2015-10-06T17:31:14.000Z
2	1001	2015-11-05T03:34:33.000Z
3	1001	2015-12-04T04:21:55.000Z
4	1001	2016-01-02T01:18:24.000Z
5	1001	2016-02-01T19:27:27.000Z
6	1001	2016-03-02T15:29:32.000Z
7	1001	2016-04-01T11:20:18.000Z
8	1001	2016-05-01T15:55:51.000Z

^MENU

Text: Formatting Best Practices

In case you want to test any of the ideas below, I have embedded the SQL environment at the bottom of this page.

Formatting Your Queries

Capitalization

You may have noticed that we have been capitalizing SELECT and FROM, while we leave table and column names lowercase. This is a common formatting convention. It is common practice to capitalize commands (SELECT, FROM), and keep everything else in your query lowercase. This makes queries easier to read, which will matter more as you write more complex queries. For now, it is just a good habit to start getting into.

Avoid Spaces in Table and Variable Names

It is common to use underscores and avoid spaces in column names. It is a bit annoying to work with spaces in SQL. In Postgres if you have spaces in column or table names, you need to refer to these columns/tables with double quotes around them (Ex: FROM "Table Name" as opposed to FROM table_name). In other environments, you might see this as square brackets instead (Ex: FROM [Table Name]).

Use White Space in Queries

SQL queries ignore spaces, so you can add as many spaces and blank lines between code as you want, and the queries are the same. This query

```
SELECT account_id FROM orders
```

is equivalent to this query:

```
SELECT account_id  
FROM orders
```

and this query (but please don't ever write queries like this):

```
SELECT account_id  
FROM orders
```

SQL isn't Case Sensitive

If you have programmed in other languages, you might be familiar with programming languages that get very upset if you do not type the correct characters in terms of lower and uppercase. SQL is not case sensitive. [The following query](#):

```
SELECT account_id  
FROM orders
```

is the same as:

```
select account_id  
from orders
```

which is also the same as:

```
SeLect account_id  
From orders
```

However, I would again urge you to follow the conventions outlined earlier in terms of fully capitalizing the commands, while leaving other pieces of your code in lowercase.

Semicolons

Depending on your SQL environment, your query may need a semicolon at the end to execute. Other environments are more flexible in terms of this being a "requirement." It is considered best practices to put a semicolon at the end of each statement, which also allows you to run multiple commands at once if your environment is able to show multiple results at once.

Best practice:

```
SELECT account_id  
FROM orders;
```

Since, our environment here doesn't require it, you will see solutions written without the semicolon:

```
SELECT account_id  
FROM orders
```

LIMIT

We have already seen the **SELECT** (to choose columns) and **FROM** (to choose tables) statements. The **LIMIT** statement is useful when you want to see just the first few rows of a table. This can be much faster for loading than if we load the entire dataset.

The **LIMIT** command is always the very last part of a query. **An example of showing just the first 10 rows of the orders table with all of the columns might look like the following:**

The screenshot shows a database query interface with the following details:

Input:

- Schema dropdown menu: accounts, orders, region, sales_reps, web_events.
- Query text:

```
1 SELECT *
2 FROM orders
3 LIMIT 10;
```
- Status: Success!
- EVALUATE button.

Output: 10 results

	id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1	1001		2015-10-06T17:31:14.000Z	123	22	24
2	1001		2015-11-05T03:34:33.000Z	190	41	57
3	1001		2015-12-04T04:21:55.000Z	85	47	0
4	1001		2016-01-02T01:18:24.000Z	144	32	0
5	1001		2016-02-01T19:27:27.000Z	108	29	28

- ⊕ We could also change the number of rows by changing the 10 to any other number of rows.
- ⊕ If there are not many rows in the table, simply reduce the rows in the table.

Quiz: LIMIT

Can You Use LIMIT?

Try it yourself below by writing a query that limits the response to only the first 15 rows and includes the `occurred_at`, `account_id`, and `channel` fields in the `web_events` table.

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1 SELECT occurred_at, account_id, channel
2 FROM web_events
3 LIMIT 15;
```

Success!

EVALUATE

Output 15 results

occurred_at	account_id	channel
2015-10-06T17:13:58.000Z	1001	direct
2015-11-05T03:08:26.000Z	1001	direct
2015-12-04T03:57:24.000Z	1001	direct
2016-01-02T00:55:03.000Z	1001	direct
2016-02-01T19:02:33.000Z	1001	direct

ORDER BY

The **ORDER BY** statement allows us to order our table by any row. If you are familiar with Excel, this is similar to the sorting you can do with filters.

The **ORDER BY** statement is always after the **SELECT** and **FROM** statements, but it is before the **LIMIT** statement. As you learn additional commands, the order of these statements will matter more. If we are using the **LIMIT** statement, it will always appear last.

Pro Tip

Remember **DESC** can be added after the column in your **ORDER BY** statement to sort in descending order, as the default is to sort in ascending order.

For example, Suppose we are in the finance department in Burch and Bosey. We want to find the most recent orders so that we can send invoices to customers.

The words **ORDER BY** will help us To complete this, by allowing us to sort orders by date.

Let's look at how data is arranged by default. We'll notice that this table appears as if it was sorted by account ID by default. Starts from 1001 and starts to grow from here.

The screenshot shows a SQL query editor interface. At the top, there are tabs for 'Personal' and 'Untitled Report'. Below the tabs, there are buttons for 'Run', 'SQL', 'Display Table', and 'Format SQL'. The SQL code in the editor is:

```
1 SELECT *
2 FROM demo.orders
3 LIMIT 1000
```

Below the code, there is a status message: "✓ Succeeded in 558ms". The results section shows a table with the following data:

		id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty	total	standard_amt_usd	gloss_amt_usd	poster_amt_usd	total_amt_usd
1	1	1001	2015-10-06 17:31:14	123	22	24	169	613.77	164.78	194.88	973.43	
2	2	1001	2015-11-05 03:34:33	190	41	57	288	948.1	307.09	462.84	1718.03	
3	3	1001	2015-12-04 04:21:55	85	47	0	132	424.15	352.03	0	776.18	
4	4	1001	2016-01-02 01:18:24	144	32	0	176	718.56	239.68	0	958.24	
5	5	1001	2016-02-01 19:27:27	108	29	28	165	538.92	217.21	227.36	983.49	
6	6	1001	2016-03-02 15:29:32	103	24	46	173	513.97	179.76	373.52	1067.25	
7	7	1001	2016-04-01 11:20:18	101	33	92	226	503.99	247.17	747.04	1498.2	

Let's ORDER BY clause to reorder the results based on the data the order was placed, which you can see in the occurred_at column. You have to write the clauses in this order, or the query will not run.

By default, order by goes from A – Z, lowest to highest, or earliest to latest, if working with dates.

The screenshot shows a SQL query editor interface with a dark theme. At the top, there are tabs for 'Personal' and 'Untitled Report'. Below the tabs are buttons for 'Run', 'Limit 100', 'Format SQL', and 'View History...'. On the left, there is a sidebar with a 'Q1' icon. The main area contains a code editor with the following SQL query:

```

1 SELECT *
2 FROM demo.orders
3 ORDER BY occurred_at
4 LIMIT 1000

```

Below the code editor, a message box indicates 'Succeeded in 631ms'. At the bottom, there is a table with the following columns: id, account_id, occurred_at, standard_qty, gloss_qty, poster_qty, total, standard_amt_usd, gloss_amt_usd, poster_amt_usd, and total_amt_usd. The table displays 1000 rows of data.

	id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty	total	standard_amt_usd	gloss_amt_usd	poster_amt_usd	total_amt_usd
1	5786	2861	2013-12-04 04:22:44	0	48	33	81	0	359.52	267.96	627.48
2	2415	2861	2013-12-04 04:45:54	490	15	11	516	2445.1	112.35	89.32	2646.77
3	4108	4311	2013-12-04 04:53:25	528	10	0	538	2634.72	74.9	0	2709.62
4	4489	1281	2013-12-05 20:29:16	0	37	0	37	0	277.13	0	277.13
5	287	1281	2013-12-05 20:33:56	492	73	0	565	2455.08	546.77	0	3001.85
6	1946	2481	2013-12-06 02:13:20	502	4	33	539	2504.98	29.96	267.96	2802.9
7	6197	3431	2013-12-06 12:55:22	53	559	315	927	264.47	4186.91	2557.8	7009.18

If you want to order the other way, you can add **desc**, short for descending, to the end of the ORDER BY clause. This will get us to the data set that we're after, which shows the 10 most recent orders, with the most recent at the top.

The screenshot shows a SQL query editor interface with a dark theme, similar to the one above. At the top, there are tabs for 'Personal' and 'Untitled Report'. Below the tabs are buttons for 'Run', 'Limit 100', 'Format SQL', and 'View History...'. On the left, there is a sidebar with a 'Q1' icon. The main area contains a code editor with the following SQL query:

```

1 SELECT *
2 FROM demo.orders
3 ORDER BY occurred_at DESC
4 LIMIT 1000

```

Below the code editor, a message box indicates 'Succeeded in 627ms'. At the bottom, there is a table with the same columns as the previous screenshot. The table displays 1000 rows of data, ordered from most recent to least recent.

	id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty	total	standard_amt_usd	gloss_amt_usd	poster_amt_usd	total_amt_usd
1	6451	3841	2017-01-02 00:02:40	42	506	302	850	209.58	3789.94	2452.24	6451.76
2	3546	3841	2017-01-01 23:50:16	291	36	26	353	1452.09	269.64	211.12	1932.85
3	6454	3861	2017-01-01 22:29:50	38	167	51	256	189.62	1250.83	414.12	1854.57
4	3554	3861	2017-01-01 22:17:26	497	0	23	520	2480.03	0	186.76	2666.79
5	6556	4051	2017-01-01 21:04:25	0	65	50	115	0	486.85	406	892.85
6	3745	4051	2017-01-01 20:52:23	495	15	0	510	2470.05	112.35	0	2582.4
7	1092	1761	2017-01-01 17:34:10	62	28	124	214	309.38	209.72	1006.88	1525.98

Quiz: ORDER BY

Practice

In order to gain some practice using **ORDER BY**:

1. Write a query to return the 10 earliest orders in the **orders** table. Include the **id**, **occurred_at**, and **total_amt_usd**.

The screenshot shows a database query interface with the following details:

Input:

- SCHEMA dropdown menu showing options: accounts, orders, region, sales_reps, web_events.
- Query code:

```
1 SELECT id,occurred_at,total_amt_usd
2 FROM orders
3 ORDER BY occurred_at
4 LIMIT 10;
```
- Status message: Success!
- EVALUATE button.

Output: 10 results

id	occurred_at	total_amt_usd
5786	2013-12-04T04:22:44.000Z	627.48
2415	2013-12-04T04:45:54.000Z	2646.77
4108	2013-12-04T04:53:25.000Z	2709.62
4489	2013-12-05T20:29:16.000Z	277.13
287	2013-12-05T20:33:56.000Z	3001.85
1946	2013-12-06T02:13:20.000Z	2802.90
6197	2013-12-06T12:55:22.000Z	7009.18
3122	2013-12-06T12:57:41.000Z	1992.13

MENU

2. Write a query to return the top 5 **orders** in terms of largest **total_amt_usd**.

The screenshot shows a data analysis interface with the following components:

- Input:** A sidebar titled "SCHEMA" containing dropdown menus for "accounts", "orders", "region", "sales_reps", and "web_events".
- Query Editor:** A code editor window containing the following SQL query:

```
1 SELECT id,occurred_at,total_amt_usd
2 FROM orders
3 ORDER BY total_amt_usd DESC
4 LIMIT 5;
```
- Status:** A message "Success!" displayed below the query.
- Evaluate Button:** A blue button labeled "EVALUATE" located next to the status message.
- Output:** A table titled "Output" showing 5 results with columns "id", "occurred_at", and "total_amt_usd". The data is as follows:

id	occurred_at	total_amt_usd
4016	2016-12-26T08:53:24.000Z	232207.07
3892	2016-06-24T13:32:55.000Z	112875.18
3963	2015-03-30T00:05:30.000Z	107533.55
5791	2014-10-24T12:06:22.000Z	95005.82
3778	2016-07-17T14:50:43.000Z	93547.84

3. Write a query to return the bottom 20 **orders** in terms of least **total**. Include the **id**, **account_id**, and **total**.

Input

SCHEMA

accounts

orders

region

sales_reps

web_events

1 SELECT id,account_id,total
2 FROM orders
3 ORDER BY total
4 LIMIT 20;

Success!

EVALUATE

Output 20 results

id	account_id	total
6375	3651	0
6435	3801	0
5001	1791	0
6323	3551	0
6312	3541	0
6281	3491	0
4770	1521	0
4490	1281	0

^MENU

Quiz: ORDER BY Part II

You will notice that many of the queries that you write in this first lesson might not produce exactly what you would hope, but by the end of all the lessons, you should have the skills to better tackle these problems.

Notice that the **ORDERS** table is quite large, so additional columns are in the right of the output you immediately see.

Questions

1. Write a query that returns the top 5 rows from **orders** ordered according to newest to oldest, but with the largest **total_amt_usd** for each date listed first for each date. You will notice each of these dates shows up as unique because of the time element. When you learn about truncating dates in a later lesson, you will better be able to tackle this question on a day, month, or yearly basis.

The screenshot shows a database query interface. On the left, there's a sidebar titled "Input" with a "SCHEMA" dropdown set to "orders". Below it are dropdowns for "accounts", "region", "sales_reps", and "web_events". On the right, the query editor contains the following SQL code:

```
1 SELECT *
2 FROM orders
3 ORDER BY occurred_at desc , total_amt_usd desc
4 LIMIT 5;
```

The status bar at the bottom of the editor says "Success!" and has a blue "EVALUATE" button. Below the editor is a table titled "Output" with 5 results. The table has columns: id, account_id, occurred_at, standard_qty, gloss_qty, and poster_qty. The data is as follows:

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
6451	3841	2017-01-02T00:02:40.000Z	42	506	302
3546	3841	2017-01-01T23:50:16.000Z	291	36	26
6454	3861	2017-01-01T22:29:50.000Z	38	167	51
3554	3861	2017-01-01T22:17:26.000Z	497	0	23
6556	4051	2017-01-01T21:04:25.000Z	0	65	50

2. Write a query that returns the top 10 rows from **orders** ordered according to oldest to newest, but with the smallest **total_amt_usd** for each date listed first for each date. **You will notice each of these dates shows up as unique because of the time element. When you learn about truncating dates in a later lesson, you will better be able to tackle this question on a day, month, or yearly basis.**

Input

HISTORY ▾ MENU ▾

SCHEMA		
accounts	▼	1 SELECT *
orders	▼	2 FROM orders
region	▼	3 ORDER BY occurred_at, total_amt_usd
sales_reps	▼	4 LIMIT 10;
web_events	▼	

Success! **EVALUATE**

Output 10 results

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
5786	2861	2013-12-04T04:22:44.000Z	0	48	33
2415	2861	2013-12-04T04:45:54.000Z	490	15	11
4108	4311	2013-12-04T04:53:25.000Z	528	10	0
4489	1281	2013-12-05T20:29:16.000Z	0	37	0
287	1281	2013-12-05T20:33:56.000Z	492	73	0

WHERE

Using the **WHERE** statement, we can subset out tables based on conditions that must be met. In the next sections, you will learn some of the common operators that are used with the **WHERE** statement.

Common symbols used within WHERE statements include:

1. **>** (greater than)
2. **<** (less than)
3. **>=** (greater than or equal to)
4. **<=** (less than or equal to)
5. **=** (equal to)
6. **!=** (not equal to)

Example :

Assume that an account manager at parch & posey is about to go to one of the most important clients.

- ⊕ We can use WHERE to create a list of all purchases made by that particular customer. It also allows us to filter a set of results based on certain criteria.
- ⊕ WHERE comes after FROM but before ORDER BY & LIMIT.
- ⊕ Phrases must be in the correct order and the query will be returned by mistake.

we'll write a query to show only orders from our top customer, which is represented here by account ID 4251. we'll write this simple equation where account ID is equal to 4251. This will produce a result set that includes all rows for which the value in account ID is equal to 4251.

The screenshot shows a SQL query editor interface. On the left, there's a sidebar titled "Input" with a "SCHEMA" dropdown menu containing "accounts", "orders", "region", "sales_reps", and "web_events". The main area displays a SQL query:

```
1 SELECT *
2 FROM orders
3 WHERE account_id = 4251
4 ORDER BY occurred_at
5 LIMIT 1000
```

The status bar at the bottom right says "Success!" and has a blue "EVALUATE" button. Below the query, the "Output" section shows a table with 13 results:

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
6719	4251	2016-06-05T01:16:37.000Z	0	78	0
4009	4251	2016-06-05T01:36:42.000Z	626	15	0
4010	4251	2016-07-04T12:34:49.000Z	498	6	2
4011	4251	2016-08-02T00:53:28.000Z	679	36	5
6720	4251	2016-08-02T01:13:08.000Z	9	0	19
4012	4251	2016-09-01T02:32:51.000Z	503	13	32

When using SQL entire rows of data are preserved intact. If you write a WHERE clause filters based on values in one column as we have done here you'll limit the result in all columns to rows that satisfy the condition.

The idea is that each row is one data point or observation and all the information contained in that row belongs together.

All the information in one row is about one order of paper.

Quiz: WHERE

Questions

Write a query that

1. Pull the first 5 rows and all columns from the **orders** table that have a dollar amount of **gloss_amt_usd** greater than or equal to 1000.

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1  SELECT *
2  FROM orders
3  WHERE gloss_amt_usd >= 1000
4  LIMIT 5;
```

HISTORY ▾ MENU ▾

Success! EVALUATE

Output 5 results

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
14	1001	2016-10-26T20:31:30.000Z	97	143	54
62	1091	2014-10-13T12:12:55.000Z	146	196	3
88	1101	2015-06-24T13:08:15.000Z	182	339	17
121	1131	2016-08-10T23:47:41.000Z	273	134	0
129	1141	2016-12-21T15:52:58.000Z	143	1045	2157

2. Pull the first 10 rows and all columns from the **orders** table that have a **total_amt_usd** less than 500.

The screenshot shows a database interface with the following details:

- Input:** A sidebar titled "SCHEMA" lists tables: accounts, orders, region, sales_reps, and web_events. The "orders" table is currently selected.
- Query:** The main area contains the following SQL code:


```

1  SELECT *
2  FROM orders
3  WHERE total_amt_usd < 500
4  LIMIT 10;
      
```
- Status:** Below the query, a green message says "Success!"
- EVALUATE:** A blue button labeled "EVALUATE" is visible.
- Output:** A table titled "Output" shows 10 results with the following data:

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
67	1091	2015-04-07T13:29:20.000Z	95	0	0
96	1101	2016-03-15T11:36:03.000Z	14	8	16
119	1131	2016-06-12T12:29:45.000Z	0	30	23
124	1131	2016-11-07T05:10:56.000Z	0	0	0
254	1251	2014-11-01T02:15:24.000Z	0	0	17
328	1291	2015-08-03T08:35:23.000Z	0	19	21
542	1421	2015-11-13T09:07:09.000Z	0	64	0
683	1501	2016-04-14T23:59:50.000Z	0	15	16
713	1521	2014-11-23T16:04:03.000Z	0	8	10
730	1521	2016-05-06T02:34:48.000Z	0	0	2
- Bottom Bar:** A dark bar with the text " MENU ".

You will notice when using these **WHERE** statements, we do not need to **ORDER BY** unless we want to actually order our data. Our condition will work without having to do any sorting of the data.

WHERE with Non-Numeric Data

The **WHERE** statement can also be used with non-numerical data. We can use the `=` and `!=` operators here. You also need to be sure to use single quotes (just be careful if you have quotes in the original text) with the text data.

Commonly when we are using **WHERE** with non-numeric data fields, we use the **LIKE**, **NOT**, or **IN** operators. We will see those before the end of this lesson!

Comparison operators can work with non-numerical data as well. Equals and not equal to make perfect sets. They allow you to select rows that match or don't match any value.

Let's look at an example from Parch & Posey accounts table which simply lists all of their customers and tell attributes about them as you see here.

The screenshot shows a database interface with the following details:

- Input Schema:** The user has selected the "accounts" schema from a dropdown menu.
- Query:** The user has entered the following SQL code:

```
1  SELECT *
2  FROM accounts
```
- Output:** The results section shows "351 results". Below it is a table with columns: id, name, website, and lat. The table contains the following data:

id	name	website	lat
1001	Walmart	www.walmart.com	40.23849561
1011	Exxon Mobil	www.exxonmobil.com	41.16915630
1021	Apple	www.apple.com	42.29049481
1031	Berkshire Hathaway	www.berkshirehathaway.com	40.94902131
1041	McKesson	www.mckesson.com	42.21709326
1051	UnitedHealth Group	www.unitedhealthgroup.com	40.08792542

There are some important rules when using these operators. If you're using an operator with values that are non-numeric, you need to put the value in single quotes. As you can see, this data set has been filtered down to the single row that matches the condition in our WHERE clause.

The screenshot shows a database query interface. In the 'Input' section, the 'SCHEMA' dropdown is set to 'accounts'. The query entered is:

```
1 SELECT *
2 FROM accounts
3 WHERE name = 'united technologies'
```

The results section shows '0 results' and an empty table with columns: id, name, website, lat, long, primary_poc, sales_rep_id.

If you change the operator to not equals, the results will show every row except the one for united technologies

The screenshot shows a database query interface. In the 'Input' section, the 'SCHEMA' dropdown is set to 'accounts'. The query entered is:

```
1 SELECT *
2 FROM accounts
3 WHERE name != 'united technologies'
```

The results section shows '351 results' and a table with columns: id, name, website, lat, long, primary_poc, sales_rep_id. The table data is as follows:

id	name	website	lat	long	primary_poc	sales_rep_id
1001	Walmart	www.walmart.com	40.23849561			
1011	Exxon Mobil	www.exxonmobil.com	41.16915630			
1021	Apple	www.apple.com	42.29049481			
1031	Berkshire Hathaway	www.berkshirehathaway.com	40.94902131			

Quiz: WHERE with Non-Numeric

Practice Question Using WHERE with Non-Numeric Data

1. Filter the accounts table to include the company name, website, and the primary point of contact (primary_poc) for Exxon Mobil in the accounts table.

The screenshot shows a database query interface with the following details:

Input:

- SCHEMA dropdown: accounts (selected)
- Code area:

```
1 SELECT name, website, primary_poc
2 FROM accounts
3 WHERE name = 'Exxon Mobil';
```
- Status message: Success!
- EVALUATE button

Output: 1 results

name	website	primary_poc
Exxon Mobil	www.exxonmobil.com	Sung Shields

Arithmetic Operators

Derived Columns

Creating a new column that is a combination of existing columns is known as a **derived** column.

Common operators include:

1. `*` (Multiplication)
2. `+` (Addition)
3. `-` (Subtraction)
4. `/` (Division)

Order of Operations

The following two statements have very different end results:

1. `Standard_qty / standard_qty + gloss_qty + poster_qty`
2. `standard_qty / (standard_qty + gloss_qty + poster_qty)`

It is likely the case you mean to calculate the statement in part 2.

Example:

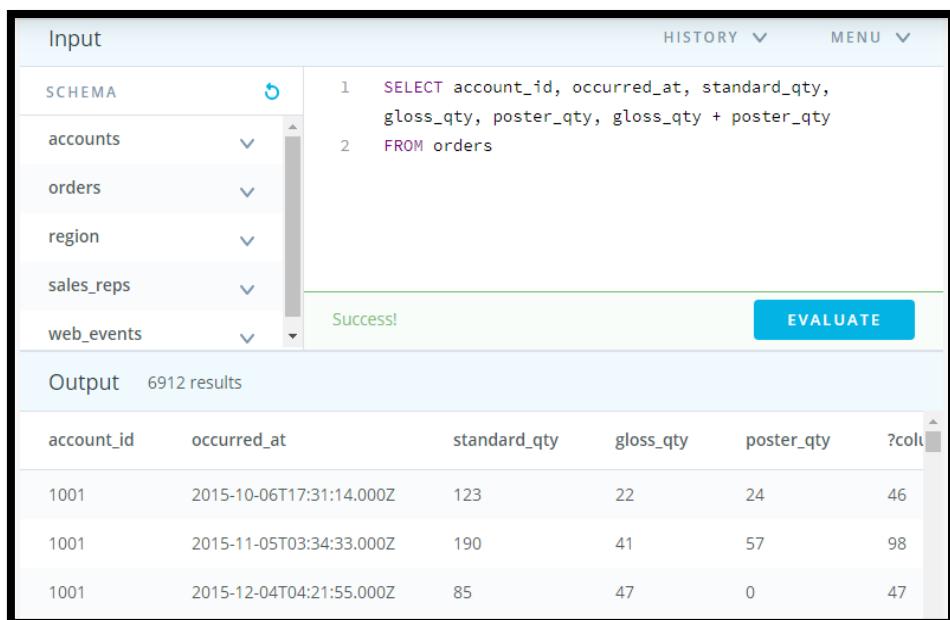
Before we calculate how much non standard paper was sold, let's check out the order quantities in each column of the orders table.

The screenshot shows a database interface with the following details:

- Input:** A sidebar titled "SCHEMA" lists tables: accounts, orders, region, sales_reps, and web_events.
- Query:** Two numbered lines of SQL code:
 1. `SELECT account_id, occurred_at, standard_qty, gloss_qty, poster_qty`
 2. `FROM orders`
- Output:** A table titled "Output" showing 6912 results with columns: account_id, occurred_at, standard_qty, gloss_qty, and poster_qty.
- Status:** The message "Success!" is displayed at the bottom of the output area.
- Buttons:** A blue "EVALUATE" button is located at the bottom right of the output area.

account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1001	2015-10-06T17:31:14.000Z	123	22	24
1001	2015-11-05T03:34:33.000Z	190	41	57
1001	2015-12-04T04:21:55.000Z	85	47	0

Now let's create a new column that adds poster and gloss paper together to create a field for non standard paper. To do this, we'll create a new line in the query and use an arithmetic operator.



The screenshot shows a data processing interface with the following components:

- Input Schema:** A sidebar on the left listing tables: accounts, orders, region, sales_reps, and web_events. The "accounts" table is currently selected.
- Query Editor:** A code editor on the right containing the following SQL query:

```
1 SELECT account_id, occurred_at, standard_qty,
2 gloss_qty, poster_qty, gloss_qty + poster_qty
3 FROM orders
```
- Status:** Below the query editor, it says "Success!"
- Evaluate Button:** A blue button labeled "EVALUATE" located at the bottom right of the status area.
- Output:** A table below the status area showing the results of the query. It has 6 columns: account_id, occurred_at, standard_qty, gloss_qty, poster_qty, and ?coll. The data shows three rows for account_id 1001.

account_id	occurred_at	standard_qty	gloss_qty	poster_qty	?coll
1001	2015-10-06T17:31:14.000Z	123	22	24	46
1001	2015-11-05T03:34:33.000Z	190	41	57	98
1001	2015-12-04T04:21:55.000Z	85	47	0	47

You'll notice that the values in this new column are equal to the sum of the gloss and poster quantities on a row-by-row basis.

In the first row, 22 gloss and 24 poster sum to 46 nonstandard in this new column we've created.

There's a special name for this. It's a derived column. Derived columns can include simple arithmetic or any number of advance calculations. What makes them special is the fact that they're not just directly selected from the underlying data, they're manipulated in some way.

Rather than stick with this ugly column name, let's add something more descriptive using an Alias. We can do this by adding AS to the end of the line that produces the derived column and then giving it a name.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

1 SELECT account_id, occurred_at, standard_qty,
gloss_qty, poster_qty, gloss_qty + poster_qty AS
nonstandard_qty
2 FROM orders

Success!

EVALUATE

Output 6912 results

account_id	occurred_at	standard_qty	gloss_qty	poster_qty	nonstandard_qty
1001	2015-10-06T17:31:14.000Z	123	22	24	46
1001	2015-11-05T03:34:33.000Z	190	41	57	98
1001	2015-12-04T04:21:55.000Z	85	47	0	47
1001	2016-01-02T01:18:24.000Z	144	32	0	32
1001	2016-02-01T19:27:27.000Z	108	29	28	57

The best practice here is to use names that are clearly descriptive, follow existing conventions, and don't include capital letters or space. We'll use nonstandard quantity.

Quiz: Arithmetic Operators

Questions using Arithmetic Operations

Using the `orders` table:

1. Create a column that divides the `standard_amt_usd` by the `standard_qty` to find the unit price for standard paper for each order. Limit the results to the first 10 orders, and include the `id` and `account_id` fields.

The screenshot shows a database query interface with the following details:

Input Schema: accounts, orders, region, sales_reps, web_events

Query:

```
1  SELECT id,account_id,standard_amt_usd/standard_qty AS
2    Unit_pric
3   FROM orders
4   LIMIT 10;
```

Output: 10 results

id	account_id	unit_pric
1	1001	4.990000000000000
2	1001	4.990000000000000
3	1001	4.990000000000000
4	1001	4.990000000000000
5	1001	4.990000000000000
6	1001	4.990000000000000
7	1001	4.990000000000000
8	1001	4.990000000000000
9	1001	4.990000000000000
10	1001	4.990000000000000

Status: Success!

EVALUATE button

2. Write a query that finds the percentage of revenue that comes from poster paper for each order. You will need to use only the columns that end with `_usd`. (Try to do this without using the total column). Include the `id` and `account_id` fields. **NOTE - you will be thrown an error with the correct solution to this question. This is for a division by zero. You will learn how to get a solution without an error to this query when you learn about CASE statements in a later section. For now, you might just add some very small value to your denominator as a work around.**

The screenshot shows a database query editor interface. On the left, there is a schema browser with tables: accounts, orders, region, sales_reps, and web_events. The 'accounts' table is currently selected. On the right, the query editor displays the following SQL code:

```

1  SELECT
    id, account_id, poster_amt_usd/(standard_amt_usd+gloss_amt_usd+poster_amt_usd+0.1) AS poster_per
2  FROM orders
  
```

The status bar at the bottom of the editor says "Success!" and there is a blue "EVALUATE" button. Below the editor, the results are displayed in a table titled "Output" with 6912 results:

id	account_id	poster_per
1	1001	0.20017873100983020554
2	1001	0.26938590211450821533
3	1001	0.000000000000000000000000000000
4	1001	0.000000000000000000000000000000
5	1001	0.23115322441261094562

Notice, the above operators combine information across columns for the same row. If you want to combine values of a particular column, across multiple rows, we will do this with aggregations. We will get to that before the end of the course!

Introduction to Logical Operators

Introduction to Logical Operators

In the next concepts, you will be learning about **Logical Operators**.

Logical Operators include:

1. **LIKE**

This allows you to perform operations similar to using **WHERE** and **=**, but for cases when you might **not** know **exactly** what you are looking for.

2. **IN**

This allows you to perform operations similar to using **WHERE** and **=**, but for more than one condition.

3. **NOT**

This is used with **IN** and **LIKE** to select all of the rows **NOT LIKE** or **NOT IN** a certain condition.

4. **AND & BETWEEN**

These allow you to combine operations where all combined conditions must be true.

5. **OR**

This allow you to combine operations where at least one of the combined conditions must be true.

Don't worry if this doesn't make total sense right now. You will get practice with each in the next sections.

LIKE

The **LIKE** operator is extremely useful for working with text. You will use **LIKE** within a **WHERE** clause. The **LIKE** operator is frequently used with %. The % tells us that we might want any number of characters leading up to a particular set of characters or following a certain set of characters. Remember you will need to use single quotes for the text you pass to the **LIKE** operator, because of this lower and uppercase letters are not the same within the string. Searching for 'T' is not the same as searching for 't'. In other SQL environments (outside the classroom), you can use either single or double quotes.

Hopefully you are starting to get more comfortable with SQL, as we are starting to move toward operations that have more applications, but this also means we can't show you every use case. Hopefully, you can start to think about how you might use these types of applications to identify phone numbers from a certain region, or an individual where you can't quite remember the full name.

Example:

Suppose a digital marketing manager at Parch & Posey wants to determine the web-based traffic that came from any page in Google. There is a problem though.

As you can see, the referrer URLs from Google all have the same domain but often have a bunch of extra stuff tacked on. In order to write a filter that capture all web traffic from Google regardless of the stuff that's tacked onto the end, you can use the LIKE operator. But first, let's see what happens if I just use an equal sign in the WHERE clause. As you can see, we get back zero results.

The screenshot shows a database query editor interface. In the top left, it says "Input". On the left, there's a "SCHEMA" dropdown menu with several options: "accounts", "orders", "region", "sales_reps", and "web_events". Below the schema dropdown is a message in red: "relation \"web_events_full\" does not exist". To the right of this message is a blue "EVALUATE" button. In the bottom left, it says "Output". The main area is empty, showing only a large red "X" icon and the text "No results due to error".

```
1 SELECT *
2 FROM web_events_full
3 WHERE referre_url
```

Now I'll use the LIKE operator to find all rows for which the referrer URL is similar to Google but with any number of characters on either side.

You'll notice that the LIKE function requires the use of wildcards. So in this case, the wildcard we're using is a percent sign.

```
Run Limit 100 Format SQL View History...
1 SELECT *
2   FROM demo.web_events_full
3 WHERE referrer_url LIKE '%google%'

Export Copy Chart Pivot
```

	id	account_id	occurred_at	channel	
1	4396	1001	2015-10-22 14:04:00	adwords	https://www.google.com
2	4399	1001	2016-01-01 15:45:00	adwords	https://www.google.com
3	4401	1001	2016-02-07 17:44:00	adwords	https://www.google.com
4	4410	1001	2016-06-22 13:48:00	adwords	https://www.google.com
5	4416	1001	2016-09-11 17:06:00	adwords	https://www.google.com

Each of these percent signs is used to represent a character or any number of characters, whatever those characters may be. So in this case, we're looking to match referrer URL for values that start with a character or any number of characters then contain the text Google, then have any number of characters on the end.

Quiz: *LIKE*

Questions using the **LIKE** operator

Use the **accounts** table to find

1. All the companies whose names start with 'C'.

The screenshot shows a database interface with the following details:

Input:

- SCHEMA:** accounts
- Code:**

```
1  SELECT name
2  FROM accounts
3  WHERE name LIKE 'C%'
```
- Status:** Success!
- EVALUATE** button

Output: 37 results

name
CVS Health
Chevron
Costco
Cardinal Health
Citigroup

2. All companies whose names contain the string 'one' somewhere in the name.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT name
2 FROM accounts
3 WHERE name LIKE '%one%';
```

Success!

EVALUATE

Output 3 results

name

Honeywell International

INTL FCStone

AutoZone

3. All companies whose names end with 's'.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT name
2 FROM accounts
3 WHERE name LIKE '%s';
```

Success!

EVALUATE

Output 4 results

name

UPS

CHS

AES

CBS

IN

The **IN** operator is useful for working with both numeric and text columns. This operator allows you to use an **=**, but for more than one item of that particular column. We can check one, two or many column values for which we want to pull data, but all within the same query. In the upcoming concepts, you will see the **OR** operator that would also allow us to perform these tasks, but the **IN** operator is a cleaner way to write these queries.

Expert Tip

In most SQL environments, you can use single or double quotation marks - and you may NEED to use double quotation marks if you have an apostrophe within the text you are attempting to pull.

In the work spaces in the classroom, note you can include an apostrophe by putting two single quotes together. Example Macy's in our work space would be 'Macy''s'.

Example:

Suppose a digital marketing manager at Parch & Posey wants to determine the web-based traffic that came from any page in Google. There is a problem though.

The IN function will allow you to filter data based on several possible values. So, if you only want to see information for the Walmart & Apple accounts, here's how you do it.

Input

HISTORY ▾ MENU ▾

SCHEMA

- accounts
- orders
- region
- sales_reps
- web_events

```

1  SELECT *
2  FROM accounts
3  WHERE name IN ('Walmart','Apple')

```

Success!

EVALUATE

Output 2 results

	id	name	website	lat	long	primary_poc
1	1001	Walmart	www.walmart.com	40.23849561	-75.10329704	Tamara Tuma
2	1021	Apple	www.apple.com	42.29049481	-76.08400942	Jodee Lupo

As with comparison operators, IN requires single quotation marks around non-numerical data can be entered directly. Let's look at the actual orders associated with those accounts.

First, make note of the account IDs here, 1001 for Walmart and 1021 for Apple. Now we see of all the individual orders that come us for account 1001&1021.

Input

HISTORY ▾ MENU ▾

SCHEMA

- accounts
- orders
- region
- sales_reps
- web_events

```

1  SELECT *
2  FROM orders
3  WHERE account_id IN ('1001','1021')

```

Success!

EVALUATE

Output 38 results

	id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1	1	1001	2015-10-06T17:31:14.000Z	123	22	24
2	2	1001	2015-11-05T03:34:33.000Z	190	41	57
3	3	1001	2015-12-04T04:21:55.000Z	85	47	0
4	4	1001	2016-01-02T01:18:24.000Z	144	32	0

And as you can see, they're all jumbled together because there is no ORDER BY clause in the query. One thing to note here is the comma between these two values.

You can put as many values as you want inside these parentheses, but you have to put a comma in between each pair of distinct values.

Quiz: IN

Questions using IN operator

1. Use the **accounts** table to find the account **name**, **primary_poc**, and **sales_rep_id** for Walmart, Target, and Nordstrom.

The screenshot shows a database query interface with the following details:

Input:

- SCHEMA:** accounts
- Output:** 3 results

Query:

```
1  SELECT name,primary_poc,sales_rep_id
2  FROM accounts
3  WHERE name IN ('Walmart','Target','Nordstrom')
```

Output:

name	primary_poc	sales_rep_id
Walmart	Tamara Tuma	321500
Target	Luba Streett	321660
Nordstrom	Yan Crater	321820

2. Use the **web_events** table to find all information regarding individuals who were contacted via the **channel** of **organic** or **adwords**.

Input

HISTORY ▾ MENU ▾

SCHEMA	accounts	orders	region	sales_reps	web_events
	▼	▼	▼	▼	▼

```
1 SELECT *
2 FROM web_events
3 WHERE channel IN ('organic','adwords')
```

Success!

EVALUATE

Output 1858 results

id	account_id	occurred_at	channel
4395	1001	2015-10-22T05:02:47.000Z	organic
4396	1001	2015-10-22T14:04:20.000Z	adwords
4399	1001	2016-01-01T15:45:54.000Z	adwords
4401	1001	2016-02-07T17:44:10.000Z	adwords
4402	1001	2016-02-27T15:27:22.000Z	organic

NOT

The **NOT** operator is an extremely useful operator for working with the previous two operators we introduced: **IN** and **LIKE**. By specifying **NOT LIKE** or **NOT IN**, we can grab all of the rows that do not meet a particular criteria.

Example:

Imagine you're a sales manager at Parch & Posey. You're considering promoting two of your top sales reps into management, but you need to figure out how you're going to divvy up all of their accounts among the other sales rep. To do this, let's first take a look at all of Parch & Posey accounts. You can see here that there are 351 in total.

A screenshot of a SQL query interface. The interface has a dark theme with a header bar containing 'Personal' and 'Untitled Report'. On the right side of the header are 'Options', 'Embed', 'Share', 'View', and a search icon. Below the header are tabs for 'Query 1', 'SQL' (which is selected), 'Display Table', and a plus sign icon. Underneath these tabs are buttons for 'Run', 'Limit 100', 'Format SQL', and 'View History...'. The main area contains a code editor with the following SQL query:

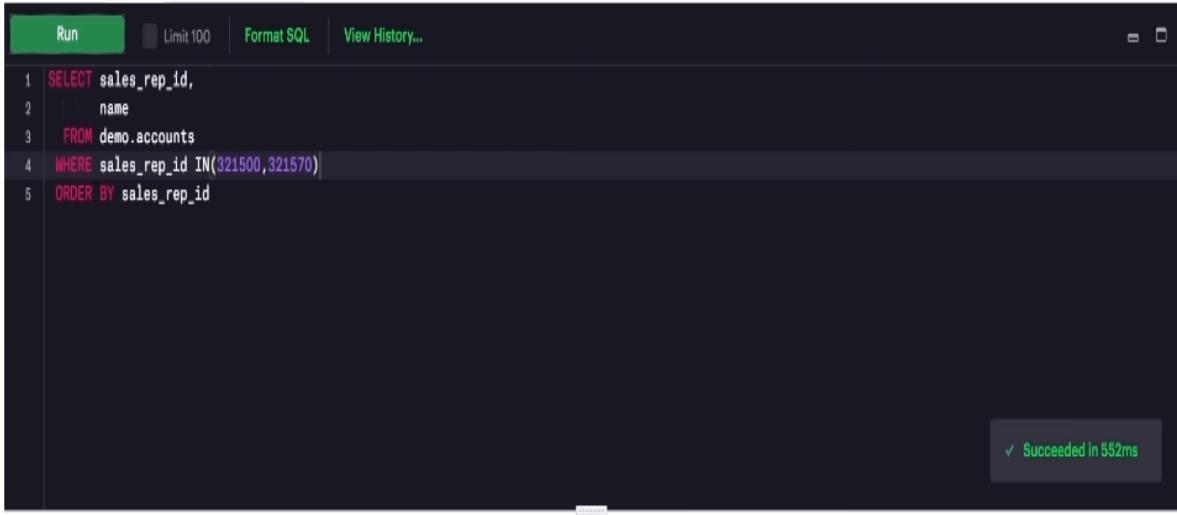
```
1 SELECT sales_rep_id,
2      name
3 FROM demo.accounts
4 ORDER BY sales_rep_id
```

At the bottom right of the code editor, there is a green 'Ready' button. Below the code editor is a table with two columns: 'sales_rep_id' and 'name'. The table contains four rows of data:

	sales_rep_id	name
1	321500	Walmart
2	321500	Freddie Mac
3	321500	Ingram Micro
4	321500	American Airlines Group

At the bottom right of the table, it says '351 rows returned'.

Now let's look at the accounts that these two high performing sales reps are working. You can see here we filtered the list of accounts to just the two reps in question, nine accounts in total.



The screenshot shows a SQL query being run in a database interface. The query is:

```
1 SELECT sales_rep_id,
2        name
3     FROM demo.accounts
4    WHERE sales_rep_id IN(321500,321570)
5 ORDER BY sales_rep_id
```

The results table has two columns: sales_rep_id and name. The data is:

	sales_rep_id	name
1	321500	Johnson Controls
2	321500	Ingram Micro
3	321500	American Airlines Group
4	321500	Walmart
5	321500	Express Scripts Holding

At the bottom, there are buttons for Export, Copy, Chart, Pivot, and a success message: ✓ Succeeded in 552ms. It also indicates 9 rows returned.

In deciding who to assign these accounts to, you might also want to consider which accounts the other sales reps are currently working. In another words, you'd like to look at all the accounts that are not listed in this query we're looking at right now.

The NOT operator will allow you to do exactly this. By simply typing NOT before this logical statement, we can turn it into the inverse.

You can see that in this new result set, none of accounts that were previously shown are appearing. Of the 351 accounts that Parch & Posey sells to, the nine in the previous query are absent from this result set, so we see 342 results.

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1 SELECT sales_rep_id, name
2 FROM accounts
3 WHERE sales_rep_id NOT IN(321500,321570)
4 ORDER BY sales_rep_id

```

Success! EVALUATE

Output 342 results

sales_rep_id	name
321510	Best Buy
321510	Delta Air Lines
321510	J.P. Morgan Chase
321510	PepsiCo

NOT also works with LIKE, the same way it works with IN. you can see here, we're looking at web traffic from Google in order to illustrate this.

```

1 SELECT *
2 FROM demo.web_events_full
3 WHERE referrer_url LIKE '%google%'

```

Ready

Export Copy Chart Pivot 1459 rows returned

	id	account_id	occurred_at	channel	referrer_url
1	4396	1001	2015-10-22 14:04:00	adwords	https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=order+paper
2	4399	1001	2016-01-01 15:45:00	adwords	https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=paper+company

By adding NOT, we can look at the inverse, all traffic from sets other than Google

The screenshot shows a SQL query being run in a database interface. The query is:

```
1 SELECT *
2 FROM demo.web_events_full
3 WHERE referrer_url NOT LIKE '%google%
```

The result of the query is displayed below, showing 2316 rows returned. The table has columns: id, account_id, occurred_at, channel, and referrer_url. Some rows are visible:

	id	account_id	occurred_at	channel	referrer_url
1	4407	1001	2016-05-07 02:28:00	banner	http://www.espn.com
2	4411	1001	2016-07-12 22:43:00	banner	http://www.espn.com
3	4413	1001	2016-07-26 21:08:00	banner	https://en.lichess.org
4	4423	1021	2015-12-26 02:15:00	banner	http://www.wellappointeddesk.com

Quiz: NOT

Questions using the NOT operator

We can pull all of the rows that were excluded from the queries in the previous two concepts with our new operator.

1. Use the **accounts** table to find the account name, primary poc, and sales rep id for all stores except Walmart, Target, and Nordstrom.

The screenshot shows a database query interface with the following details:

Input:

- SCHEMA dropdown: accounts (selected)
- Other tables: orders, region, sales_reps, web_events

```
1 SELECT name,primary_poc,sales_rep_id
2 FROM accounts
3 WHERE name NOT IN ('Walmare','Target','Nordstrom')
```

Output: 349 results

name	primary_poc	sales_rep_id
Walmart	Tamara Tuma	321500
Exxon Mobil	Sung Shields	321510
Apple	Jodee Lupo	321520
Berkshire Hathaway	Serafina Banda	321530
McKesson	Angeles Crusoe	321540

The query was successful, as indicated by the green "Success!" message next to the EVALUATE button.

2. Use the **web_events** table to find all information regarding individuals who were contacted via any method except using **organic** or **adwords** methods.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```

1  SELECT *
2  FROM web_events
3  WHERE channel NOT IN ('organic','adwords')
    
```
- Status:** Success!
- Output:** 7215 results
- Table Headers:** id, account_id, occurred_at, channel
- Table Data:**

1	1001	2015-10-06T17:13:58.000Z	direct
2	1001	2015-11-05T03:08:26.000Z	direct
3	1001	2015-12-04T03:57:24.000Z	direct
4	1001	2016-01-02T00:55:03.000Z	direct
5	1001	2016-02-01T19:02:33.000Z	direct
6	1001	2016-03-02T15:15:22.000Z	direct

Use the **accounts** table to find:

1. All the companies whose names do not start with 'C'.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```

1  SELECT *
2  FROM accounts
3  WHERE name not LIKE 'C%'
    
```
- Status:** Success!
- Output:** 314 results
- Table Headers:** id, name, website, lat
- Table Data:**

1001	Walmart	www.walmart.com	40.23849561
1011	Exxon Mobil	www.exxonmobil.com	41.16915630
1021	Apple	www.apple.com	42.29049481
1031	Berkshire Hathaway	www.berkshirehathaway.com	40.94902131
1041	McKesson	www.mckesson.com	42.21709326
1051	UnitedHealth Group	www.unitedhealthgroup.com	40.08792542

2. All companies whose names do not contain the string 'one' somewhere in the name.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```

1 SELECT *
2 FROM accounts
3 WHERE name NOT LIKE '%one%'
```
- Status:** Success!
- EVALUATE** button.
- Output:** 348 results.
- Table Headers:** id, name, website, lat.
- Table Data:** A list of company entries including Walmart, Exxon Mobil, Apple, Berkshire Hathaway, McKesson, and UnitedHealth Group.

3. All companies whose names do not end with's'.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```

1 SELECT *
2 FROM accounts
3 WHERE name NOT LIKE '%s'
```
- Status:** Success!
- EVALUATE** button.
- Output:** 347 results.
- Table Headers:** id, name, website, lat.
- Table Data:** A list of company entries including Walmart, Exxon Mobil, Apple, Berkshire Hathaway, McKesson, and UnitedHealth Group.

AND and BETWEEN

The **AND** operator is used within a **WHERE** statement to consider more than one logical clause at a time. Each time you link a new statement with an **AND**, you will need to specify the column you are interested in looking at. You may link as many statements as you would like to consider at the same time. This operator works with all of the operations we have seen so far including arithmetic operators (+, *, -, /). **LIKE**, **IN**, and **NOT** logic can also be linked together using the **AND** operator.

BETWEEN Operator

Sometimes we can make a cleaner statement using **BETWEEN** than we can using **AND**. Particularly this is true when we are using the same column for different parts of our **AND** statement. In the previous lesson, we probably should have used **BETWEEN**.

Instead of writing :

```
WHERE column >= 6 AND column <= 10
```

we can instead write, equivalently:

```
WHERE column BETWEEN 6 AND 10
```

Example:

Imagine yourself as a sales manager at Parch & Posey, you want to see which of your customers bought paper a while ago and might be due to make a purchase soon.

Let's assume it's January 2017. You'd probably want to look up customers who made purchases somewhere between three &nine months prior, or between April and October of 2016.

In order to do this, you'll need to filter based on multiple criteria using AND.

Let's dig into this example by pulling orders that took place after April 1st 2016.

The screenshot shows a SQL query editor interface. At the top, there are buttons for "Run", "Limit 100", "Format SQL", and "View History...". The main area contains the following SQL code:

```
1 SELECT *
2 FROM demo.orders
3 WHERE occurred_at >= '2016-04-01'
4 ORDER BY occurred_at
```

In the bottom right corner of the editor window, there is a green success message: "✓ Succeeded in 702ms". Below the editor, there is a table with the following columns: id, account_id, occurred_at, standard_qty, gloss_qty, poster_qty, total, standard_amt_usd, gloss_amt_usd, poster_amt_usd, and total_amt_usd. The table displays four rows of data, corresponding to the results of the executed query.

	id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty	total	standard_amt_usd	gloss_amt_usd	poster_amt_usd	total_amt_usd
1	1757	2341	2016-04-01 04:39:00	500	0	21	521	2495	0	170.52	2665.52
2	1591	2181	2016-04-01 06:10:39	298	194	0	492	1487.02	1453.06	0	2940.08
3	4312	1001	2016-04-01 11:15:27	497	618	152	1267	2480.03	4628.82	1234.24	8343.09
4	7	1001	2016-04-01 11:20:18	101	33	92	226	503.99	247.17	747.04	1498.2

Next, let's exclude orders more recent than October 1st 2016. We can add this to our WHERE clause using AND.

The screenshot shows a SQL query editor interface. On the left, there is a schema browser with tables: accounts, orders, region, sales_reps, and web_events. The "orders" table is currently selected. The main area contains the following SQL code:

```
1 SELECT *
2 FROM orders
3 WHERE occurred_at >= '2016-04-01' AND occurred_at <=
4 '2016-10-01'
5 ORDER BY occurred_at
```

In the bottom right corner of the editor window, there is a green success message: "Success!". Below the editor, there is a table titled "Output" with 1823 results. The table has columns: id, account_id, occurred_at, standard_qty, gloss_qty, and poster_qty. The table displays five rows of data, corresponding to the results of the executed query.

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1757	2341	2016-04-01T04:39:00.000Z	500	0	21
1591	2181	2016-04-01T06:10:39.000Z	298	194	0
4312	1001	2016-04-01T11:15:27.000Z	497	618	152
7	1001	2016-04-01T11:20:18.000Z	101	33	92

Let's wrap this up by reordering these result to confirm that orders after October 1st were excluded.

The screenshot shows a SQL query editor interface. In the top left, there is a schema browser with tables: accounts, orders, region, sales_reps, and web_events. The main area contains a SQL query:

```
1 SELECT *
2 FROM orders
3 WHERE occurred_at >= '2016-04-01' AND occurred_at <=
4          '2016-10-01'
      ORDER BY occurred_at DESC
```

The status bar at the bottom says "Success!" and there is a blue "EVALUATE" button. Below the query, the results are displayed in a table:

Output 1823 results					
id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
516	1411	2016-09-30T19:50:54.000Z	555	32	0
4639	1411	2016-09-30T19:36:42.000Z	0	37	4
6818	4351	2016-09-30T18:23:04.000Z	264	444	318

You can see that all orders after October 1st, 2016, have not been filtered out. One important thing to note here is that the AND operator allows you to run two complete logical statements. When this query runs, it evaluates a row to see if occurred_at is greater than or equal to April 1st, 2016. Then evaluates the same row to see if occurred_at is less than or equal to October 1st 2016. These need to be independent logical statements, even though they're both operating on the same column name.

We can't just type where occurred_at is greater than or equal to April 1st and less than or equal to October 1st. less than or equal to October 1st isn't a logical statement that can be evaluated as

The screenshot shows a SQL query editor with a dark theme. The query is identical to the one above:

```
1 SELECT *
2 FROM demo.orders
3 WHERE occurred_at >= '2016-04-01' AND <= '2016-10-01'
4 ORDER BY occurred_at DESC
```

A yellow warning box at the bottom right says "There was a problem with your query". Below the editor, an error message is displayed:

Looks like something went wrong with your query. [Get help here.](#)

syntax error at or near "<="
Position: 68

```
FROM demo.orders
WHERE occurred_at >= '2016-04-01' AND <= '2016-10-01'
^
ORDER BY occurred_at DESC
```

true or false, so the query will return an error.

Quiz: AND and BETWEEN

Questions using AND and BETWEEN operators

1. Write a query that returns all the **orders** where the **standard_qty** is over 1000, the **poster_qty** is 0, and the **gloss_qty** is 0.

Input

HISTORY ▾ MENU ▾

SCHEMA	▼
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1  SELECT *
2  FROM orders
3  WHERE standard_qty > 1000 AND poster_qty = 0 AND
      gloss_qty = 0;
4
```

Success! EVALUATE

Output 2 results

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
2613	2951	2016-08-15T00:06:12.000Z	1171	0	0
3260	3491	2014-08-29T22:43:00.000Z	1552	0	0

2. Using the **accounts** table find all the companies whose names do not start with 'C' and end with 's'.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts
- Query:**

```

1 SELECT name
2 FROM accounts
3 WHERE name NOT LIKE 'C%' AND name LIKE '%s'

```

- Status:** Success!
- Output:** 2 results
- Results:**

name
UPS

3. Use the **web_events** table to find all information regarding individuals who were contacted via **organic** or **adwords** and started their account at any point in 2016 sorted from newest to oldest.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts
- Query:**

```

1 SELECT *
2 FROM web_events
3 WHERE channel IN ('organic', 'adwords') AND occurred_at
    BETWEEN '2016-01-01' AND '2017-01-01'
4 ORDER BY occurred_at DESC;

```

- Status:** Success!
- Output:** 1025 results
- Results:**

id	account_id	occurred_at	channel
8493	4141	2016-12-31T16:31:23.000Z	organic
5661	1851	2016-12-31T06:55:38.000Z	organic
5562	1791	2016-12-31T02:08:50.000Z	adwords
7703	3351	2016-12-30T21:06:53.000Z	adwords
7921	3521	2016-12-30T20:15:48.000Z	organic
6416	2401	2016-12-30T17:51:36.000Z	adwords

OR

Similar to the **AND** operator, the **OR** operator can combine multiple statements. Each time you link a new statement with an **OR**, you will need to specify the column you are interested in looking at. You may link as many statements as you would like to consider at the same time. This operator works with all of the operations we have seen so far including arithmetic operators (+, *, -, /), **LIKE**, **IN**, **NOT**, **AND**, and **BETWEEN** logic can all be linked together using the **OR** operator.

When combining multiple of these operations, we frequently might need to use parentheses to assure that logic we want to perform is being executed correctly. The example below shows of these situations.

Example :

Imagine yourself a sales manager at Porch & Posey. You've determined that another great way to expand revenue is to sell new types of paper to existing customers. In order to figure out the best prospects, you'd like to see existing customers whose orders omitted some type of paper. Orders different types of paper are marked in three separate columns in the orders table. So you'll need to write a query that makes logical comparisons on each of those columns. You can see that most orders are for all three types of paper.

Now let's start to build in the logic to find orders where one paper type might not have been ordered. You can do that using OR.

OR is a logical operator in SQL that allows you to select rows that satisfy either of two conditions. It works similarly to AND which selects the rows that satisfy both of two conditions.

Let's build this out by first selecting only orders for which standard paper was not included.

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1   SELECT account_id, occurred_at, standard_qty,
2     gloss_qty, poster_qty
3   FROM orders
4   WHERE standard_qty = 0

```

Success!

EVALUATE

Output 825 results

account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1131	2016-06-12T12:29:45.000Z	0	30	23
1131	2016-11-07T05:10:56.000Z	0	0	0
1191	2016-12-15T20:49:47.000Z	0	171	24
1251	2014-11-01T02:15:24.000Z	0	0	17
1291	2015-08-03T08:35:23.000Z	0	19	21

You can see that produces 825 results. Now let's add another paper type using OR.

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1   SELECT account_id, occurred_at, standard_qty,
2     gloss_qty, poster_qty
3   FROM orders
4   WHERE standard_qty = 0 OR gloss_qty = 0

```

Success!

EVALUATE

Output 1750 results

account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1021	2015-12-11T16:53:18.000Z	483	0	21
1021	2016-01-10T09:29:45.000Z	535	0	34
1031	2016-12-25T03:54:27.000Z	1148	0	215
1051	2016-10-01T00:48:28.000Z	486	0	1
1081	2015-03-13T09:48:32.000Z	309	0	25
1091	2014-08-15T05:38:05.000Z	150	0	19

You can see that this result set is largest than the previous one. That's because we've made this filter more inclusive. It will now return results where either one of these two logical statements resolves is true. By adding more logical statement, will arrive at a query that returns all orders where some type of paper was omitted.

The screenshot shows a database query interface with the following details:

- Input:** A sidebar titled "SCHEMA" lists tables: accounts, orders, region, sales_reps, and web_events. The "orders" table is currently selected.
- Query Editor:** The query is:


```

1 SELECT account_id, occurred_at, standard_qty,
2       gloss_qty, poster_qty
3   FROM orders
4 WHERE standard_qty = 0 OR gloss_qty = 0 OR poster_qty =
      0
      
```
- Status:** The status bar at the bottom right says "Success!"
- EVALUATE:** A blue button to execute the query.
- Output:** A table titled "Output" showing 2557 results with columns: account_id, occurred_at, standard_qty, gloss_qty, and poster_qty.
- Data Preview:** The first few rows of the output table are:

account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1001	2015-12-04T04:21:55.000Z	85	47	0
1001	2016-01-02T01:18:24.000Z	144	32	0
1001	2016-07-30T03:26:30.000Z	101	36	0
1011	2016-12-21T10:59:34.000Z	527	14	0
1021	2015-10-12T02:21:56.000Z	516	23	0
1021	2015-11-11T07:37:01.000Z	497	61	0

OR can be combined with other operators for using parentheses. Going back to our example where we were looking for orders that did not contain all three types of paper, we can make that even more targeted. The best prospects might be folks who omitted some type of paper, and also ordered recently. Let's keep our filters from that previous query, but also require that the order have occurred after October 1st 2016.

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1  SELECT account_id, occurred_at, standard_qty,
2      gloss_qty, poster_qty
3  FROM orders
4  WHERE standard_qty = 0 OR gloss_qty = 0 OR poster_qty =
0 AND occurred_at >= '2016-10-01'

```

Success! EVALUATE

Output 1900 results

account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1011	2016-12-21T10:59:34.000Z	527	14	0
1021	2015-12-11T16:53:18.000Z	483	0	21
1021	2016-01-10T09:29:45.000Z	535	0	34
1031	2016-12-25T03:54:27.000Z	1148	0	215
1041	2016-11-13T10:11:52.000Z	307	22	0
1051	2016-10-01T00:48:28.000Z	486	0	1

You'll notice that this time we wrap the **OR** statements in parentheses. You can think of the parentheses as turning three logical statements into one big logical statement, that will resolve to true or false. As long as one of these three statements is true, the entire block in the parentheses is considered to be true. The next logical statement after **AND**, will also be evaluated as true or false. And the query will only return rows for which the entire set of **OR** statements is true, and the data filter is true. Looking at this first row, you can see that "**occurred_at**" is after October 1st 2016, and "**poster_qty**" is zero, fulfilling the logical statement created by the parentheses.

Quiz: OR

Questions using the OR operator

- Find list of **orders** ids where either **gloss_qty** or **poster_qty** is greater than 4000.
Only include the **id** field in the resulting table.

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1 SELECT id,gloss_qty,poster_qty
2 FROM orders
3 WHERE (gloss_qty > 4000 OR poster_qty > 4000)
```

Success!

EVALUATE

Output 14 results

id	gloss_qty	poster_qty
362	23	11380
731	12012	31
1191	32	9301
1913	6450	45
1939	0	4078
3778	3	11226
3858	28	9989
3963	14281	0
4016	16	28262
4230	4489	2
4698	484	4901
4942	10744	95
5791	10	11691
6590	488	11034

MENU

2. Write a query that returns a list of **orders** where the **standard_qty** is zero and either the **gloss_qty** or **poster_qty** is over 1000.

The screenshot shows a database query interface with the following details:

Input Schema:

- accounts
- orders
- region
- sales_reps
- web_events

Query:

```

1 SELECT *
2 FROM orders
3 WHERE standard_qty = 0 AND (gloss_qty > 1000 OR
    poster_qty > 1000)
  
```

Output: 17 results

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1913	2461	2013-12-29T09:50:38.000Z	0	6450	45
4369	1111	2015-11-15T17:47:46.000Z	0	486	2988
4391	1161	2016-06-04T08:58:10.000Z	0	106	2967
4420	1191	2016-05-21T23:21:14.000Z	0	43	1448
4448	1231	2016-06-25T12:27:15.000Z	0	3178	23
4698	1451	2015-02-26T06:13:21.000Z	0	484	4901

3. Find all the company names that start with a 'C' or 'W', and the primary contact **contains** 'ana' or 'Ana', but it doesn't contain 'eana'.

The screenshot shows a database query interface with the following details:

Input Schema:

- accounts
- orders
- region
- sales_reps
- web_events

Query:

```

1 SELECT *
2 FROM accounts
3 WHERE (name LIKE 'C%' OR name LIKE 'W%') AND
    ((primary_poc LIKE '%ana%' OR primary_poc LIKE '%Ana%')
    AND primary_poc NOT LIKE '%eana%');
  
```

Output: 2 results

id	name	website	lat	long	prim
1061	CVS Health	www.cvshealth.com	41.46779585	-73.76763638	Anabe
1361	Comcast	www.comcastcorporation.com	42.54154764	-76.24992387	Shana

Recap & Looking Ahead

Commands

You have already learned a lot about writing code in SQL! Let's take a moment to recap all that we have covered before moving on:

Statement	How to Use It	Other Details
SELECT	<code>SELECT Col1, Col2, ...</code>	Provide the columns you want
FROM	<code>FROM Table</code>	Provide the table where the columns exist
LIMIT	<code>LIMIT 10</code>	Limits based number of rows returned
ORDER BY	<code>ORDER BY Col</code>	Orders table based on the column. Used with DESC .
WHERE	<code>WHERE Col > 5</code>	A conditional statement to filter your results
LIKE	<code>WHERE Col LIKE "%me%"</code>	Only pulls rows where column has 'me' within the text
IN	<code>WHERE Col IN ('Y', 'N')</code>	A filter for only rows with column of 'Y' or 'N'
NOT	<code>WHERE Col NOT IN ('Y', 'N')</code>	NOT is frequently used with LIKE and IN
AND	<code>WHERE Col1 > 5 AND Col2 < 3</code>	Filter rows where two or more conditions must be true
OR	<code>WHERE Col1 > 5 OR Col2 < 3</code>	Filter rows where at least one condition must be true
BETWEEN	<code>WHERE Col BETWEEN 3 AND 5</code>	Often easier syntax than using an AND

Other Tips

Though SQL is **not case sensitive** (it doesn't care if you write your statements as all uppercase or lowercase), we discussed some best practices. **The order of the key words does matter!** Using what you know so far, **you will want to write your statements as:**

```
SELECT col1, col2  
FROM table1  
WHERE col3 > 5 AND col4 LIKE '%os%'  
ORDER BY col5  
LIMIT 10;
```

Notice, you can retrieve different columns than those being used in the **ORDER BY** and **WHERE** statements. Assuming all of these column names existed in this way (col1, col2, col3, col4, col5) within a table called table1, this query would run just fine.

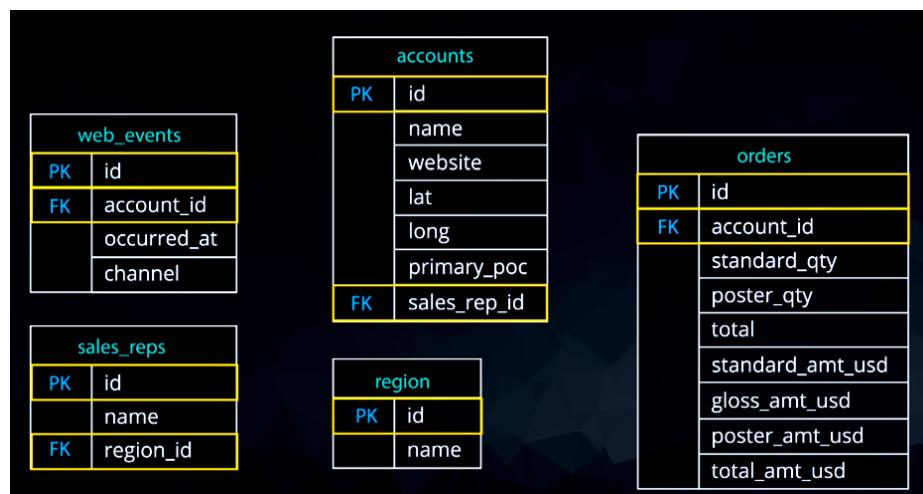
Looking Ahead

In the next lesson, you will be learning about **JOINs**. This is the real secret (well not really a secret) behind the success of SQL as a language. **JOINs** allow us to combine multiple tables together. All of the operations we learned here will still be important moving forward, but we will be able to answer much more complex questions by combining information from multiple tables! You have already mastered so much - potentially writing your first code ever, but it is about to get so much better!

Motivation

Up to this point, we've only been working with one table at a time. But the real power of SQL comes from working with data across multiple tables at once.

The term relational database refers to the fact that tables within it relate to one another. They contain common identifiers that allow information from multiple tables to be easily combined.



Keeping all of the company's data, from purchasing transactions, to employee job satisfaction, to inventory in a single Excel dataset doesn't make a ton of sense. The table would hold a ton of information and it'd be hard to determine a row call and structure for so many different types of data.

Databases give us the flexibility to keep things organized neatly in their own tables, so that they're easy to find and work with while also allowing us to combine tables as needed to solve problems that require several types of data.

In this lesson, we're going to take a look at how to leverage SQL to link tables together. You're about to see why SQL is one of the most popular environments for working with data as we learn who to write what are known as joints.

Why Would We Want to Split Data Into Separate Tables?

To understand what JOINs are and why they're helpful, let's think about Parch & Posey's orders table. Looking back at this table, you'll notice that none of the orders say the name of the customer.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events
- Query:** SELECT * FROM orders
- Status:** Success!
- Output:** 6912 results
- Table Headers:** id, account_id, occurred_at, standard_qty, gloss_qty, poster_qty
- Table Data:** (5 rows shown)

1	1001	2015-10-06T17:31:14.000Z	123	22	24
2	1001	2015-11-05T03:34:33.000Z	190	41	57
3	1001	2015-12-04T04:21:55.000Z	85	47	0
4	1001	2016-01-02T01:18:24.000Z	144	32	0
5	1001	2016-02-01T19:27:27.000Z	108	29	28

Instead, the table refers to customers by numerical values in the **account_id** column, we'll need to JOIN another table in order to connect this data to names. But first, why isn't the customer's name in this table in the first place?

There are a few reasons that someone might have made the decision to separate orders from the information about the customers placing those orders.

There are several reasons why relation databases are like this.

let's focus on two of the most important once:

first, orders and accounts are different types of objects and will be easier to organize if kept separate.

Second, this multi-table structure allows queries to execute more quickly.

let's look at each of these points more closely. The accounts and orders tables store fundamentally different types of objects.

Parch & Posey probably only wants one account per company, and they wanted to be up-to-date with the latest information.



Orders, on the other hand, are likely to stay the same once they are entered, and certainly, once they're filled. A given customer might have multiple orders, and rather than change a past order, Parch & Posey might just add a new one. Because these objects operate differently, it makes sense for them to live in different tables.



Another reason accounts and orders might be stored separately has to do with database can modify data. When you write a query, its execution speed depends on the amount of data you're asking the database to read and the number and type of calculations you're asking it to make.

Imagine a world where the account names and addresses are added into the orders table.

This means, the table would have six additional columns:

One for account name, and five that make up the street address, city, state, zip code, and country.

Let's say, customer change address. In this world, the five address columns need to be update retroactively for every single order. This means five updates times the number of orders that they've made.

id	account_id	time	Name	Address	City	State	Zip	Country
1	1001	time1	name1	add1	c1	s1	z1	count1
2	1001	time2	name1	add1	c1	s1	z1	count1
3	1011	time3	name2	add2	c2	s2	z2	count2
4	1011	time4	name2	add2	c2	s2	z2	count2
5	1001	time5	name1	add1	c1	s1	z1	count1

By contrast, keeping account details in a separate table makes this only total of five updates. The larger the data set, the more that this matters.

Database Normalization

When creating a database, it is really important to think about how data will be stored. This is known as **normalization**, and it is a huge part of most SQL classes. If you are in charge of setting up a new database, it is important to have a thorough understanding of database **normalization**.

There are essentially three ideas that are aimed at database normalization:

1. Are the tables storing logical groupings of the data?
2. Can I make changes in a single location, rather than in many tables for the same information?
3. Can I access and manipulate data quickly and efficiently?

However, most analysts are working with a database that was already set up with the necessary properties in place. As analysts of data, you don't really need to think too much about data **normalization**. You just need to be able to pull the data from the database, so you can start drawing insights. This will be our focus in this lesson.

So, we understand that all of the information related to an account is not in the orders table, but **why not?**

There are more reasons for this kind of structure but for now, what you really need to know is how to connect the account data and order data. The mechanism for doing that is a JOIN, which you'll learn next.

Introduction to JOINS

This entire lesson will be aimed at **JOINS**. The whole goal of **JOIN** statements is to allow us to pull from more than one table at a time.

Again - **JOINS** are useful for allowing us to pull data from multiple tables. This is both simple and powerful all at the same time.

With the addition of the **JOIN** statement to our toolkit, we will also be adding the **ON** statement.

JOINS will make your queries a little bit more complicated. Let's break down all the parts of a JOIN to see what they do before you write some of your own. We'll start by writing what's called an **INNER JOIN** and we'll move on to other types of **JOINS** shortly.

In order to write a **JOIN**, the first thing we need are **SELECT** and **FROM** clauses just like any other query. In this case, we're looking to add account names to each order. So we'll start with the **orders** table. You'll notice I'm running this query even though it's not yet finished.

The screenshot shows a database query editor interface. At the top, there's a menu bar with 'HISTORY' and 'MENU'. Below the menu is a sidebar labeled 'Input' containing a 'SCHEMA' dropdown with options: accounts, orders, region, sales_reps, and web_events. To the right of the schema dropdown is a code editor window displaying the following SQL query:

```
1 SELECT *
2 FROM orders
```

The code editor has a status message 'Success!' and a blue 'EVALUATE' button. Below the code editor is an 'Output' section showing the results of the query. The output header is '6912 results' and it includes columns: id, account_id, occurred_at, standard_qty, gloss_qty, and poster_qty. The data table contains five rows of data:| | | | | | |
| --- | --- | --- | --- | --- | --- |
| | | | | | |
| 1 | 1001 | 2015-10-06T17:31:14.000Z | 123 | 22 | 24 |
| 2 | 1001 | 2015-11-05T03:34:33.000Z | 190 | 41 | 57 |
| 3 | 1001 | 2015-12-04T04:21:55.000Z | 85 | 47 | 0 |
| 4 | 1001 | 2016-01-02T01:18:24.000Z | 144 | 32 | 0 |
| 5 | 1001 | 2016-02-01T19:27:27.000Z | 108 | 29 | 28 |

This is an easy way to make sure I haven't made any typos or errors in the first part of my query. Since it worked, we can move on to write the **JOIN**.

The next thing we need is a **JOIN** clause. You can think of this like a second **FROM** clause. It identifies the table where the data that we want to join lives.

In this case, account names live in the accounts table, so that's what we've written. Finally, we need to specify the relationship between the two tables. This happens by writing a logical statement in the **ON** clause.

Now that the JOIN is complete, let's add some columns from the accounts table into our **SELECT** statement to complete the query.

A screenshot of a SQL query editor interface. At the top, there are buttons for "Run", "Limit 100", "Format SQL", and "View History...". The main area contains the following SQL code:

```
1 SELECT orders.*,
2      accounts.*
3 FROM demo.orders
4 JOIN demo.accounts
5 ON orders.account_id = accounts.id
```

In the bottom right corner of the editor window, there is a green button with a checkmark and the text "Succeeded in 1s". Below the editor, there is a toolbar with buttons for "Export", "Copy", "Chart", and "Pivot". To the right of the toolbar, it says "6912 rows returned". The bottom half of the screen shows a table with the following data:

		id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty	total	standard_amt_usd	gloss_amt_usd	poster_amt_usd	total_amt_usd		name
1	1001	1001		2015-10-06 17:31:14	123	22	24	169	613.77	164.78	194.88	973.43		Walmart
2	1001	1001		2015-11-05 03:34:33	190	41	57	288	948.1	307.09	462.84	1718.03		Walmart
3	1001	1001		2015-12-04 04:21:55	85	47	0	132	424.15	352.03	0	776.18		Walmart
4	1001	1001		2016-01-02 01:18:24	144	32	0	176	718.56	239.68	0	958.24		Walmart

You can see that the results now contain information about the orders and the accounts that placed them.

Your First JOIN

Write Your First JOIN

Below is a **JOIN**, you will get a lot of practice writing these, and there is no better way to learn than practice. You will notice, we have introduced two new parts to our regular queries: **JOIN** and **ON**. The **JOIN** introduces the second table from which you would like to pull data, and the **ON** tells you how you would like to merge the tables in the **FROM** and **JOIN** statements together.

Example:

The screenshot shows a database query interface with the following details:

Input Schema:

- accounts
- orders
- region
- sales_reps
- web_events

Query:

```
1 SELECT orders.*  
2 FROM orders  
3 JOIN accounts  
4 ON orders.account_id = accounts.id;
```

Output: 6912 results

Success!

EVALUATE button

Download CSV button

Output Data:

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1	1001	2015-10-06T17:31:14.000Z	123	22	24
2	1001	2015-11-05T03:34:33.000Z	190	41	57
3	1001	2015-12-04T04:21:55.000Z	85	47	0
4	1001	2016-01-02T01:18:24.000Z	144	32	0
5	1001	2016-02-01T19:27:27.000Z	108	29	28
6	1001	2016-03-02T15:29:32.000Z	103	24	46

What to Notice

We are able to pull data from two tables:

1. [orders](#)
2. [accounts](#)

Above, we are only pulling data from the [orders](#) table since in the SELECT statement we only reference columns from the [orders](#) table.

The [ON](#) statement holds the two columns that get linked across the two tables. This will be the focus in the next concepts.

Additional Information

If we wanted to only pull individual elements from either the [orders](#) or [accounts](#) table, we can do this by using the exact same information in the [FROM](#) and [ON](#) statements. However, in your [SELECT](#) statement, **you will need to know how to specify tables and columns in the [SELECT](#) statement:**

1. The table name is always [before](#) the period.
2. The column you want from that table is always [after](#) the period.

For example, if we want to pull only the [account name](#) and the dates in which that account placed an order, but none of the other columns, **we can do this with the following query:**

The screenshot shows a database query interface with the following details:

- Input:** A sidebar on the left lists available schemas: accounts, orders, region, sales_reps, and web_events. The 'accounts' schema is currently selected.
- Query Editor:** The main area contains the following SQL code:

```
1 SELECT accounts.name, orders.occurred_at
2 FROM orders
3 JOIN accounts
4 ON orders.account_id = accounts.id;
```
- Status:** Below the code, a green "Success!" message is displayed.
- EVALUATE:** A blue button is located at the bottom right of the code editor.
- Output:** Below the status, it says "6912 results". To the right is a "Download CSV" link.
- Results Table:** A table displays the first few rows of the output:

name	occurred_at
Walmart	2015-10-06T17:31:14.000Z
Walmart	2015-11-05T03:34:33.000Z
Walmart	2015-12-04T04:21:55.000Z
Walmart	2016-01-02T01:18:24.000Z

This query only pulls two columns, not all the information in these two tables. Alternatively, the below query pulls all the columns from both the **accounts** and **orders** table.

The screenshot shows a database interface with the following details:

- Input:** A schema browser on the left lists tables: accounts, orders, region, sales_reps, and web_events. The 'orders' table is currently selected.
- Query:** The code entered is:


```

1  SELECT *
2  FROM orders
3  JOIN accounts
4  ON orders.account_id = accounts.id;
      
```
- Status:** The message "Success!" is displayed.
- EVALUATE:** A blue button to run the query.
- Output:** 6912 results are shown in a table with columns: id, account_id, occurred_at, standard_qty, gloss_qty, and poster_qty.
- Actions:** A "Download CSV" link is available.

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1001	1001	2015-10-06T17:31:14.000Z	123	22	24
1001	1001	2015-11-05T03:34:33.000Z	190	41	57
1001	1001	2015-12-04T04:21:55.000Z	85	47	0
1001	1001	2016-01-02T01:18:24.000Z	144	32	0
1001	1001	2016-02-01T19:27:27.000Z	108	29	28
1001	1001	2016-03-02T15:29:32.000Z	103	24	46

And the first query you ran pull all the information from only the **orders** table:

The screenshot shows a database interface with the following details:

- Input:** A schema browser on the left lists tables: accounts, orders, region, sales_reps, and web_events. The 'orders' table is currently selected.
- Query:** The code entered is identical to the previous screenshot:


```

1  SELECT orders.*
2  FROM orders
3  JOIN accounts
4  ON orders.account_id = accounts.id;
      
```
- Status:** The message "Success!" is displayed.
- EVALUATE:** A blue button to run the query.
- Output:** 6912 results are shown in a table with columns: id, account_id, occurred_at, standard_qty, gloss_qty, and poster_qty.
- Actions:** A "Download CSV" link is available.

1	1001	2015-10-06T17:31:14.000Z	123	22	24
2	1001	2015-11-05T03:34:33.000Z	190	41	57
3	1001	2015-12-04T04:21:55.000Z	85	47	0
4	1001	2016-01-02T01:18:24.000Z	144	32	0
5	1001	2016-02-01T19:27:27.000Z	108	29	28
6	1001	2016-03-02T15:29:32.000Z	103	24	46

Joining tables allows you access to each of the tables in the **SELECT** statement through the table name, and the columns will always follow a **.** after the table name.

Quiz: Your First JOIN

Quiz Questions

1. Try pulling all the data from the **accounts** table, and all the data from the **orders** table.

Input

HISTORY ▾ MENU ▾

SCHEMA	▼
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1  SELECT orders.*, accounts.*  
2  FROM accounts  
3  JOIN orders  
4  ON accounts.id = orders.account_id;
```

Success! EVALUATE

Output 6912 results Download CSV

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1001	1001	2015-10-06T17:31:14.000Z	123	22	24
1001	1001	2015-11-05T03:34:33.000Z	190	41	57
1001	1001	2015-12-04T04:21:55.000Z	85	47	0
1001	1001	2016-01-02T01:18:24.000Z	144	32	0
1001	1001	2016-02-01T19:27:27.000Z	108	29	28
1001	1001	2016-03-02T15:29:32.000Z	103	24	46

2. Try pulling **standard_qty**, **gloss_qty**, and **poster_qty** from the **orders** table, and the **website** and the **primary_poc** from the **accounts** table.

Input

HISTORY ▾ MENU ▾

SCHEMA		1
accounts	▼	1 <code>SELECT orders.standard_qty, orders.gloss_qty,</code>
orders	▼	2 <code>orders.poster_qty, accounts.website,</code>
region	▼	3 <code>JOIN accounts</code>
sales_reps	▼	4 <code>ON orders.account_id = accounts.id</code>
web_events	▼	

Success! EVALUATE

Output 6912 results Download CSV

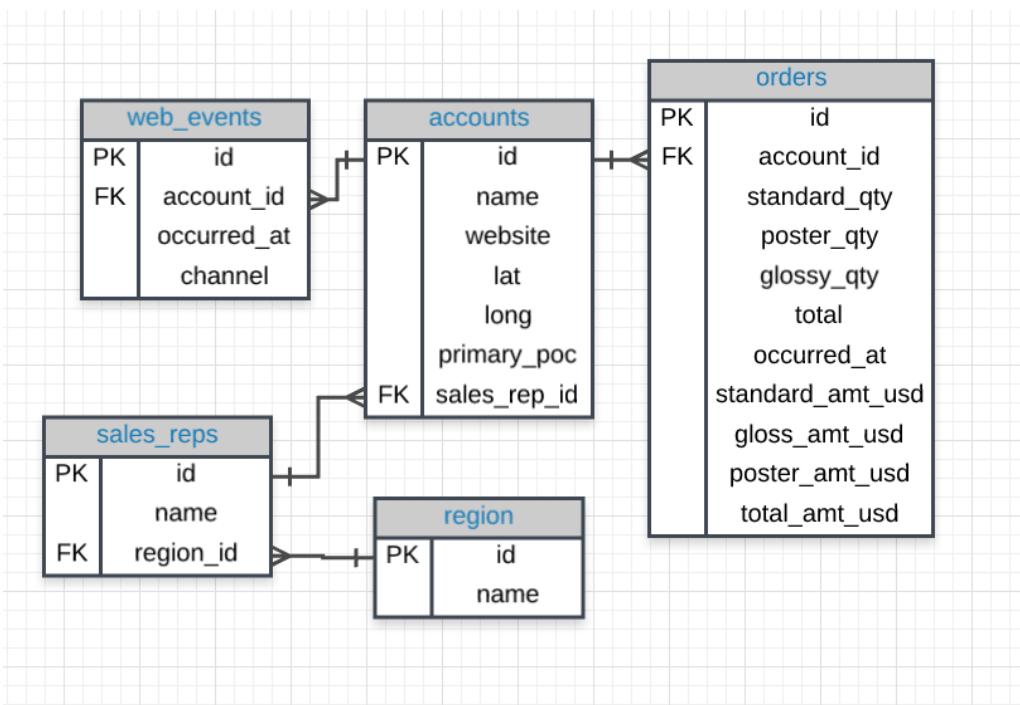
standard_qty	gloss_qty	poster_qty	website	primary_poc
123	22	24	www.walmart.com	Tamara Tum.
190	41	57	www.walmart.com	Tamara Tum.
85	47	0	www.walmart.com	Tamara Tum.
144	32	0	www.walmart.com	Tamara Tum.
108	29	28	www.walmart.com	Tamara Tum.
103	24	46	www.walmart.com	Tamara Tum.

ERD Reminder

Entity Relationship Diagrams

From the last lesson, you might remember that an entity relationship diagram (**ERD**) is a common way to view data in a database. It is also a key element to understanding how we can pull data from multiple tables.

It will be beneficial to have an idea of what the **ERD** looks like for Parch & Posey handy, so I have posted it again below.



Tables & Columns

In the Parch & Posey database there are 5 tables:

1. [web_events](#)
2. [accounts](#)
3. [orders](#)
4. [sales_reps](#)
5. [region](#)

You will notice some of the columns in the tables have **PK** or **FK** next to the column name, while other columns don't have a label at all.

If you look a little closer, you might notice that the **PK** is associated with the first column in every table. The **PK** here stands for **primary key**. **A primary key exists in every table, and it is a column that has a unique value for every row.**

If you look at the first few rows of any of the tables in our database, you will notice that this first, **PK**, column is always unique. For this database it is always called **id**, but that is not true of all databases.

Primary and Foreign Keys

Primary Key (PK)

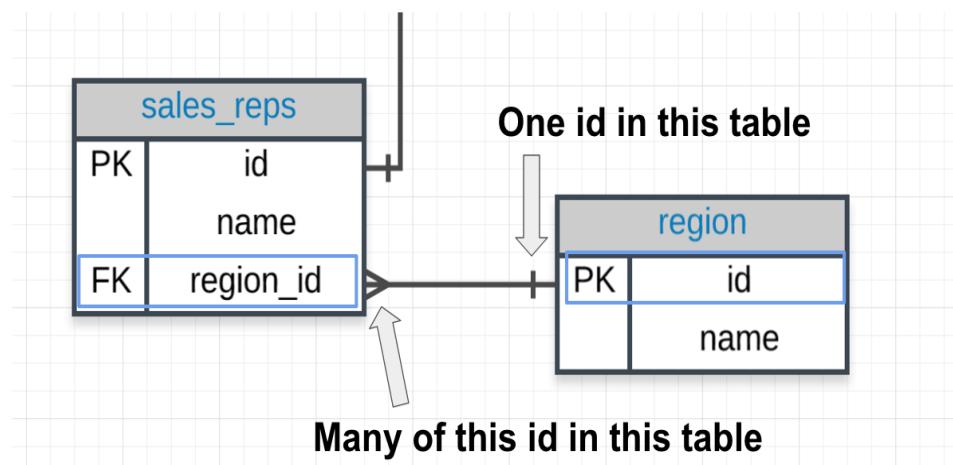
A **primary key** is a unique column in a particular table. This is the first column in each of our tables. Here, those columns are all called **id**, but that doesn't necessarily have to be the name. **It is common that the primary key is the first column in our tables in most databases.**

Foreign Key (FK)

A **foreign key** is when we see a primary key in another table. **We can see these in the previous ERD the foreign keys are provided as:**

1. region_id
2. account_id
3. sales_rep_id

Each of these is linked to the **primary key** of another table. **An example is shown in the image below:**



Primary - Foreign Key Link

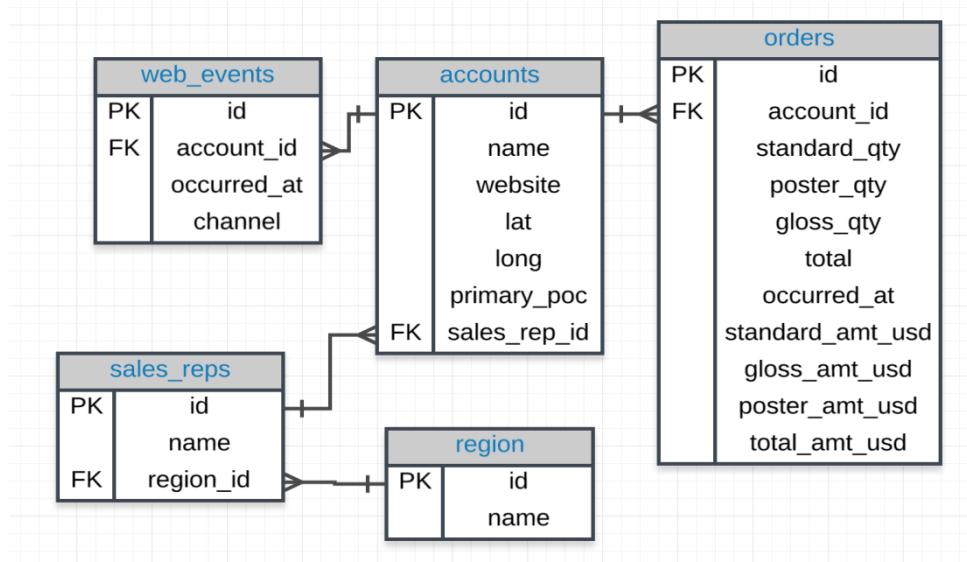
In the above image you can see that:

1. The **region_id** is the foreign key.
2. The **region_id** is **linked** to **id** - this is the primary-foreign key link that connects these two tables.
3. The crow's foot shows that the **FK** can actually appear in many rows in the **sales_reps** table.
4. While the single line is telling us that the **PK** shows that **id** appears only once per row in this table.

If you look through the rest of the database, you will notice this is always the case for a primary-foreign key relationship.

Quiz: Primary - Foreign Key Relationship

Helpful ERD For Answering the Below Questions



QUESTION 1 OF 3

Match the appropriate definition or description to each term or column.

ON `accounts.id = web_events.id`

ON `web_events.id = accounts.id`

ON `accounts.account_id = web_events.id`

DEFINITION OR COLUMN DESCRIPTION	TERM OR COLUMN
Has a unique value for every row in that column. There is one in every table.	Primary Key
The link to the primary key that exists in another table.	Foreign Key
The primary key in every table of our example database.	id
A foreign key that exists in both the web_events and orders tables.	account_id
The ON statement associated with a JOIN of the web_events and accounts tables.	ON web_events.account_id =accounts.id

QUESTION 2 OF 3

Select all that are true for primary keys.

There is one and only one of these columns in every table.

They are a column in a table.

There might be more than one primary key for a table.

They are a row in a table.

Every database only has one primary key for the whole database.

QUESTION 3 OF 3

Select all that are true of foreign keys.

They are always linked to a primary key.

They are unique for every row in a table.

Every table must have a foreign key.

A table can only have one foreign key.

In the above database, every foreign key is associated with the crow-foot notation, which suggests it can appear multiple times in the column of a table.

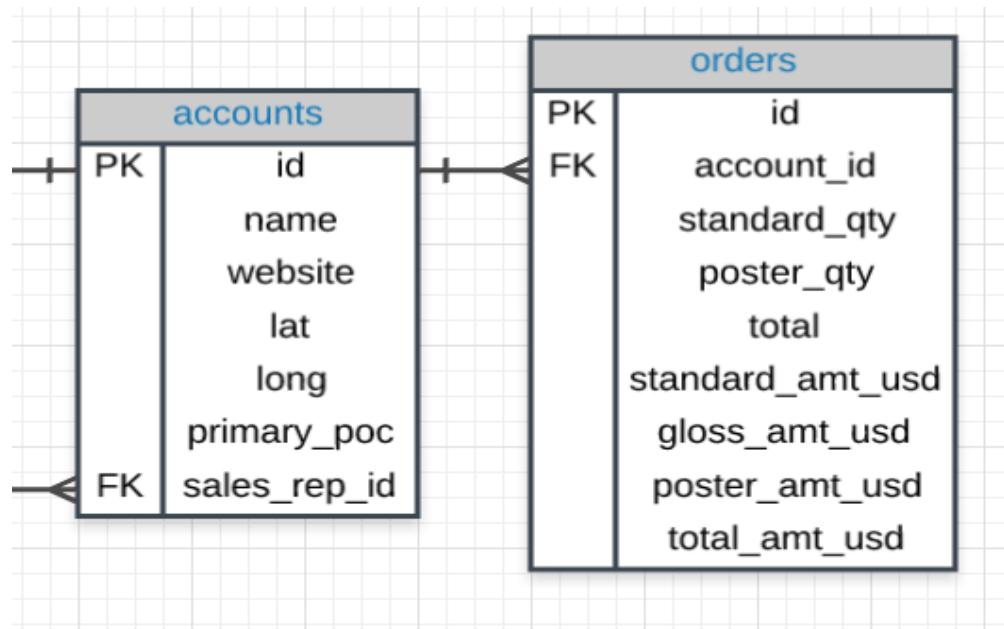
JOIN Revisited

JOIN Revisited

Let's look back at the first JOIN you wrote.

```
SELECT orders.*  
FROM orders  
JOIN accounts  
ON orders.account_id = accounts.id;
```

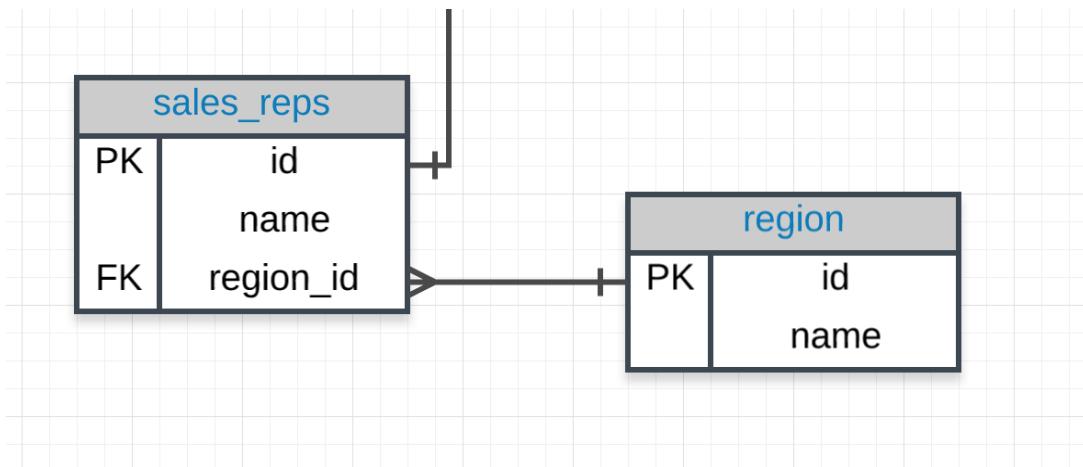
Here is the ERD for these two tables:



Notice

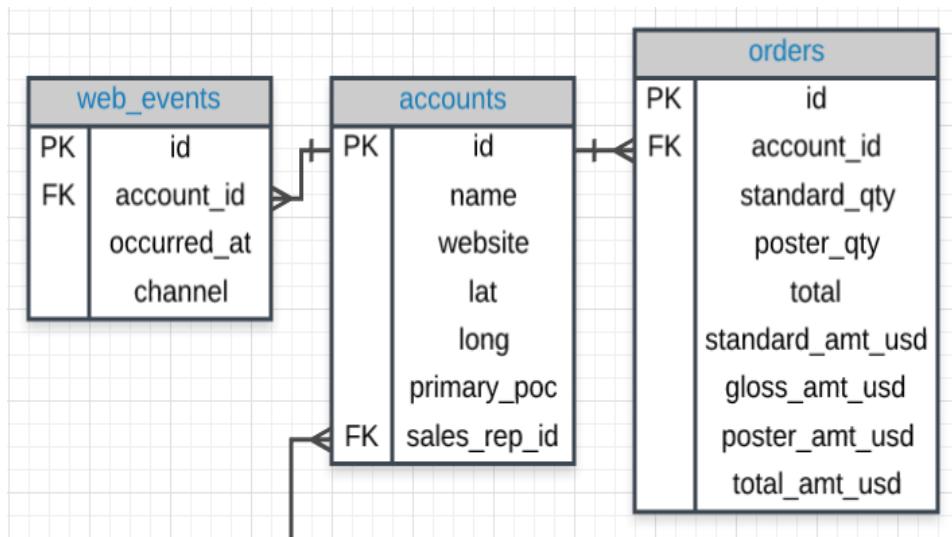
Notice our SQL query has the two tables we would like to join - one in the **FROM** and the other in the **JOIN**. Then in the **ON**, we will **ALWAYS** have the **PK** equal to the **FK**:

The way we join any two tables is in this way: linking the **PK** and **FK** (generally in an **ON** statement).



JOIN More than Two Tables

This same logic can actually assist in joining more than two tables together. Look at the three tables below.



The Code

If we wanted to join all three of these tables, we could use the same logic. The code below pulls all of the data from all of the joined tables.

```
SELECT *
FROM web_events
JOIN accounts
ON web_events.account_id = accounts.id
JOIN orders
ON accounts.id = orders.account_id
```

Alternatively, we can create a **SELECT** statement that could pull specific columns from any of the three tables. Again, our **JOIN** holds a table, and **ON** is a link for our **PK** to equal the **FK**.

To pull specific columns, the **SELECT** statement will need to specify the table that you are wishing to pull the column from, as well as the column name. We could pull only three columns in the above by changing the select statement to the below, **but maintaining the rest of the JOIN information:**

```
SELECT web_events.channel, accounts.name, orders.total
```

We could continue this same process to link all of the tables if we wanted. For efficiency reasons, we probably don't want to do this unless we actually need information from all of the tables.

Quiz: JOIN Revisited

QUIZ QUESTION

Practice

Use the image above to assist you. If we wanted to join the `sales_reps` and `region` tables together, how would you do it

- ON `sales_reps.id` = `region.id`

- ON `sales_reps.id` = `region.name`

- ON `sales_reps.region_id` = `region.id`

- ON `region.id` = `sales_reps.id`

Alias

When we **JOIN** tables together, it is nice to give each table an **alias**. Frequently an alias is just the first letter of the table name. You actually saw something similar for column names in the **Arithmetic Operators** concept.

Example:

```
FROM table name AS t1  
JOIN tablename2 AS t2
```

Before, you saw something like:

```
SELECT col1 + col2 AS total, col3
```

Frequently, you might also see these statements without the **AS** statement. Each of the above could be written in the following way instead, and they would still produce the **exact same results**:

```
FROM table name t1  
JOIN tablename2 t2
```

and

```
SELECT col1 + col2 total, col3
```

Aliases for Columns in Resulting Table

While aliasing tables is the most common use case. It can also be used to alias the columns selected to have the resulting table reflect a more readable name.

Example:

```
Select t1.column1 alias name, t2.column2 aliasname2  
FROM table name AS t1  
JOIN tablename2 AS t2
```

The alias name fields will be what shows up in the returned table instead of t1.column1 and t2.column2

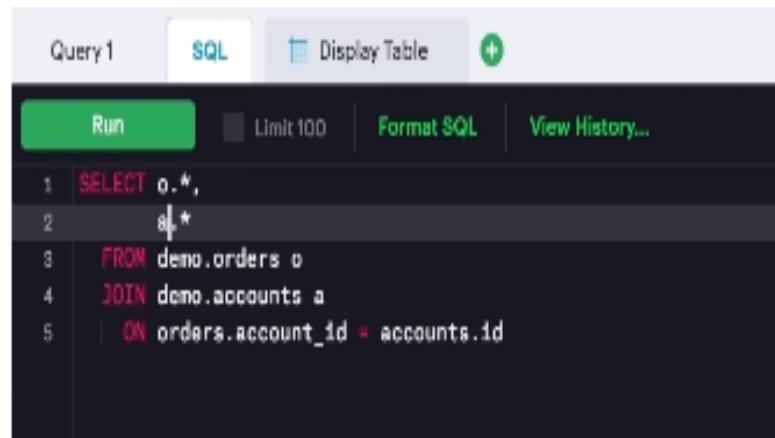
aliasname	aliasname2
example row	example row
example row	example row

When performing joins, it's easiest to give your table names aliases. Orders is a pretty long and annoying name to type. O is much easier. You can give a table an alias by adding a space after the table name and typing the intended name of the alias. As with column names, the best practice here is to use all lower case letters and underscores instead of spaces.

A screenshot of a SQL query editor interface. The top bar has tabs for 'Query 1', 'SQL' (which is selected), 'Display Table', and a plus icon. Below the tabs are buttons for 'Run', 'Limit 100', 'Format SQL', and 'View History...'. The main area contains a code editor with the following SQL query:

```
1 SELECT orders.*,  
2      accounts.*  
3 FROM demo.orders o  
4 JOIN demo.accounts a  
5 ON orders.account_id = accounts.id
```

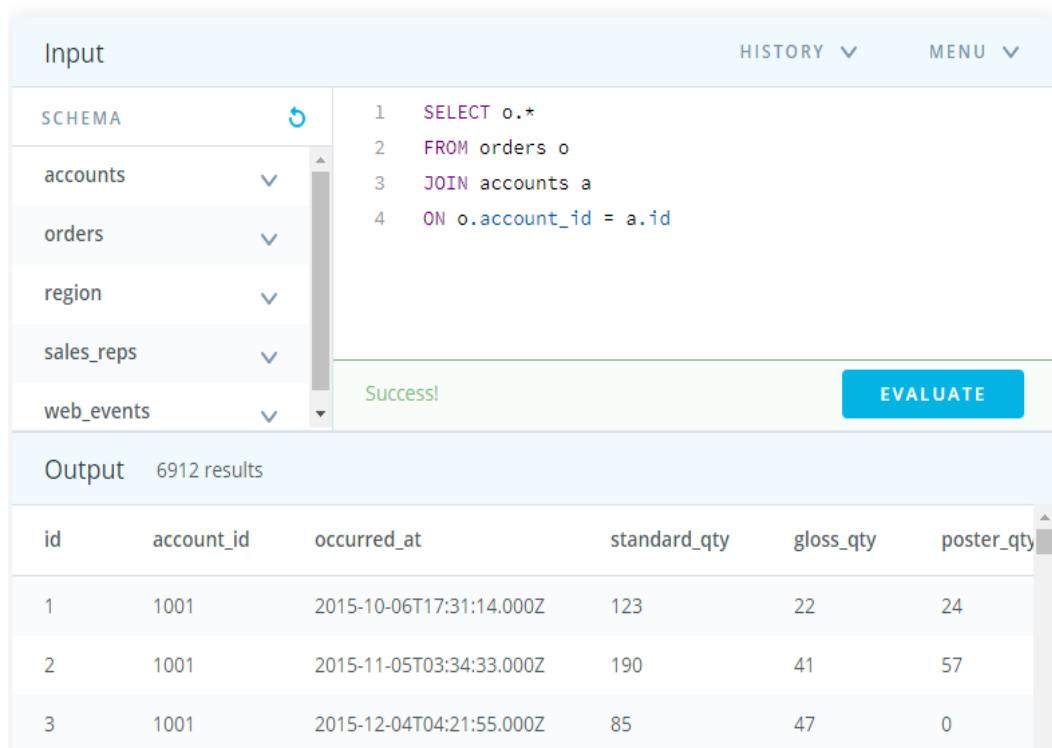
Now that we've given our tables aliases, we can refer to columns in the SELECT clause by using those alias names. We can also do the same thing anywhere else in the query that columns from these tables appear.



The screenshot shows a SQL editor interface with a tab labeled "Query 1". The "SQL" tab is selected. Below the tabs, there are buttons for "Run", "Limit 100", "Format SQL", and "View History...". The main area contains the following SQL code:

```
1 | SELECT o.*,
2 |     a.*
3 | FROM demo.orders o
4 | JOIN demo.accounts a
5 | ON orders.account_id = accounts.id
```

In this case, the only other place the ON clause. As you can see, the results are exactly the same as the previous query without the aliases.



The screenshot shows a database interface with a sidebar titled "Input" containing a "SCHEMA" dropdown with options: accounts, orders, region, sales_reps, and web_events. To the right of the schema dropdown is a list of four numbered SQL statements:

- 1 | SELECT o.*
- 2 | FROM orders o
- 3 | JOIN accounts a
- 4 | ON o.account_id = a.id

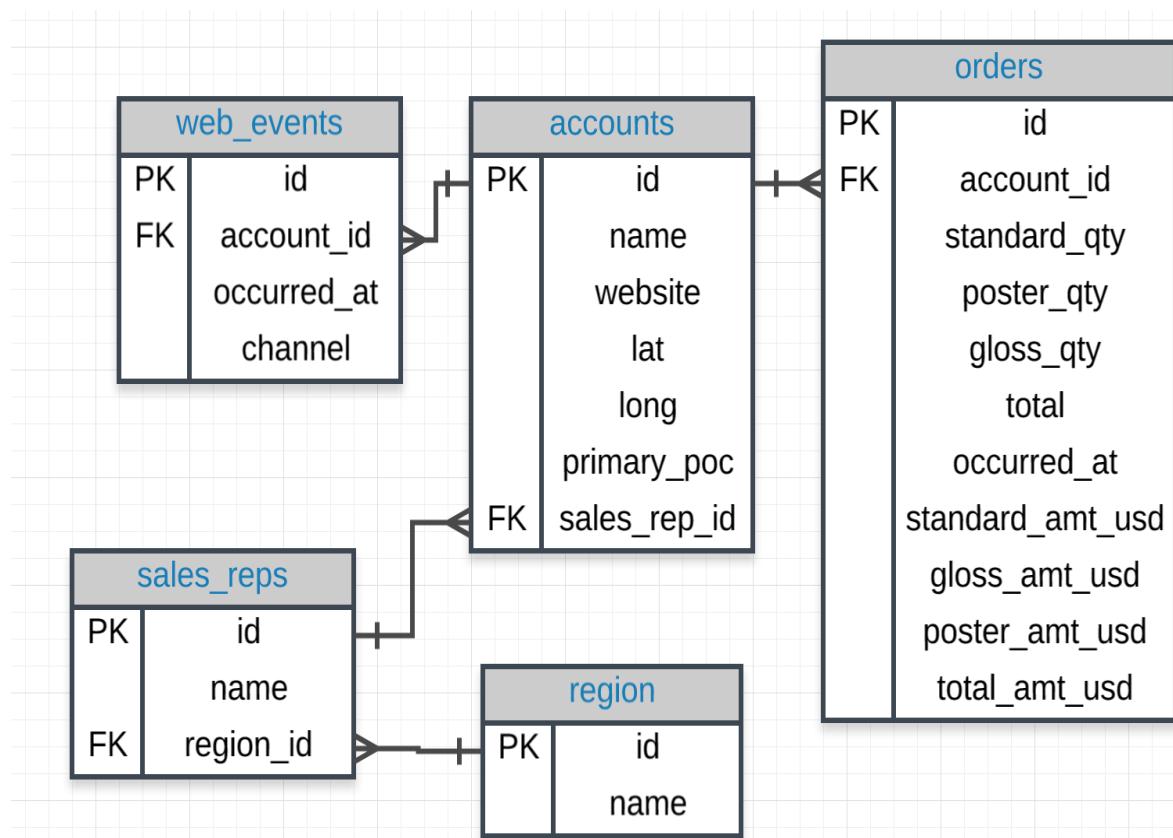
Below the schema dropdown, the word "Success!" is displayed next to a blue "EVALUATE" button. At the bottom of the interface, there is an "Output" section showing "6912 results" and a table with columns: id, account_id, occurred_at, standard_qty, gloss_qty, and poster_qty. The table contains three rows of data:

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1	1001	2015-10-06T17:31:14.000Z	123	22	24
2	1001	2015-11-05T03:34:33.000Z	190	41	57
3	1001	2015-12-04T04:21:55.000Z	85	47	0

Quiz: JOIN Questions Part I

Question Mania

Now that you have been introduced to JOINS, let's practice to build your skills and comfort with this new tool. Below I have provided the **ERD** and a bunch of questions. The solutions for the questions can be found on the next concept for you to check your answers or just in case you get stuck!



Questions

1. Provide a table for all **web_events** associated with account **name** of **Walmart**. There should be three columns. Be sure to include the **primary_poc**, time of the event, and the **channel** for each event. Additionally, you might choose to add a fourth column to assure only **Walmart** events were chosen.

Input

HISTORY ▾ MENU ▾

SCHEMA	▼
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1
2 SELECT a.primary_poc, w.occurred_at, w.channel, a.name
3 FROM web_events w
4 JOIN accounts a
5 ON w.account_id = a.id
6 WHERE a.name = 'Walmart';
```

Success! EVALUATE

Output 39 results

primary_poc	occurred_at	channel	name
Tamara Tuma	2015-10-06T17:13:58.000Z	direct	Walmart
Tamara Tuma	2015-11-05T03:08:26.000Z	direct	Walmart
Tamara Tuma	2015-12-04T03:57:24.000Z	direct	Walmart
Tamara Tuma	2016-01-02T00:55:03.000Z	direct	Walmart
Tamara Tuma	2016-02-01T19:02:33.000Z	direct	Walmart

2. Provide a table that provides the **region** for each **sales_rep** along with their associated **accounts**. Your final table should include three columns: the region **name**, the sales rep **name**, and the account **name**. Sort the accounts alphabetically (A-Z) according to account name.

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1  SELECT r.name region, s.name rep, a.name account
2  FROM sales_reps s
3  JOIN region r
4  ON s.region_id = r.id
5  JOIN accounts a
6  ON a.sales_rep_id = s.id
7  ORDER BY a.name;
```

Success!

EVALUATE

Output 351 results

region	rep	account
Northeast	Sibyl Lauria	3M
Midwest	Chau Rowles	Abbott Laboratories
Midwest	Julie Starr	AbbVie
Southeast	Earlie Schleusner	ADP
West	Marquetta Laycock	Advance Auto Parts

3. Provide the **name** for each region for every **order**, as well as the account **name** and the **unit price** they paid (**total_amt_usd/total**) for the order. Your final table should have 3 columns: **region name**, **account name**, and **unit price**. A few accounts have 0 for **total**, so I divided by (**total + 0.01**) to assure not dividing by zero.

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1  SELECT r.name region, a.name account,
2    o.total_amt_usd/(o.total + 0.01) unit_price
3    FROM region r
4    JOIN sales_reps s
5    ON s.region_id = r.id
6    JOIN accounts a
7    ON a.sales_rep_id = s.id
8    JOIN orders o
9    ON o.account_id = a.id;

```

Success! **EVALUATE**

Output 6912 results

region	account	unit_price
Northeast	Walmart	5.7596000236672386
Northeast	Walmart	5.9651748203187389
Northeast	Walmart	5.8797060828725097
Northeast	Walmart	5.4442361229475598
Northeast	Walmart	5.9601842312587116

QUESTION 1 OF 2

Mark all the below that are true.

- The **ON** statement **should** always occur with the foreign key being equal to the primary key.
 - SQL requires** the **ON** statement always occur with the foreign key being equal to the primary key.
-
- JOIN** statements allow us to pull data from multiple tables in a **SQL** database.
-
- You can use all of the commands we saw in the first lesson along with **JOIN** statements.
-
- The only way to pull data together from multiple tables is using a **JOIN** statement.

QUESTION 2 OF 2

Select all of the below statements that are true.

- If we want to alias a table, we **must** use an **AS** in our query.
 - If we want to alias a column, we **must** use an **AS** in our query.
-
- Aliasing is common to shorten table names when we start **JOINing** multiple tables together.

Motivation for Other JOINS

Expert Tip

You have had a bit of an introduction to these **one-to-one** and **one-to-many** relationships when we introduced **PKs** and **FKs**. Notice, traditional databases do not allow for **many-to-many** relationships, as these break the schema down pretty quickly. The types of relationships that exist in a database matter less to analysts, but you do need to understand why you would perform different types of **JOINS**, and what data you are pulling from the database.

One thing you might have noticed when working with inner joins is that the data you join can have a one to one or a one to many relationship. In other ward, one account might only have placed a single order while another account might have placed many orders.

Since we're focused primarily on the orders, you probably missed a different kind of relationship. A few of these accounts have just developed brand new relationships with Parch & Posey sales team and haven't placed any orders yet. So, when we join these tables, what happens to the accounts without any orders? Where do they go?

It turns out you might want to keep them around or drop them from your result set depending on what you're going for.

For example, if you're simply attaching account names to each order, then excluding these accounts without orders is probably fine. That's exactly what we did in our inner join earlier.

If your goal is to count up all of the accounts in the region along with their quantities of paper purchased, you might want to include the accounts without any orders. This is done using an outer join.

The next section, I'll show you a few different types of joins and give the opportunity to practice writing some yourself.

LEFT and RIGHT JOINS

JOINS

Notice each of these new **JOIN** statements pulls all the same rows as an **INNER JOIN**, which you saw by just using **JOIN**, but they also potentially pull some additional rows.

If there is not matching information in the **JOINED** table, then you will have columns with empty cells. These empty cells introduce a new data type called **NULL**. You will learn about **NULLs** in detail in the next lesson, but for now you have a quick introduction as you can consider any cell without data as **NULL**.

For the following example, I'll use small subsets of the accounts and orders tables shown here.

ACCOUNTS		ORDERS		
id	name	id	account_id	total
1001	Walmart	1	1001	169
1011	Exxon Mobil	2	1001	288
1021	Apple	17	1011	541
		18	1021	539
		19	1021	558
		24	1031	1363

First, let's quickly recap the inner join. This will return only rows that appear in both tables.

Writing the following INNER JOIN query to join the accounts and orders tables will pull these columns from the accounts table and these column in the orders table, only for the values of the account_id in the orders table that match with the id field in the accounts table.

ACCOUNTS

id	name
1001	Walmart
1011	Exxon Mobil
1021	Apple

ORDERS

id	account_id	total
1	1001	169
2	1001	288
17	1011	541
18	1021	539
19	1021	558
24	1031	1363

SELECT a.id, a.name, o.total
FROM orders o
JOIN accounts a
ON orders.account_id = accounts.id

This will pull these three rows from the accounts table and match them with these five rows from the orders table.

ACCOUNTS

id	name
1001	Walmart
1011	Exxon Mobil
1021	Apple

ORDERS

id	account_id	total
1	1001	169
2	1001	288
17	1011	541
18	1021	539
19	1021	558
24	1031	1363

So the table we'll get back will look like this.

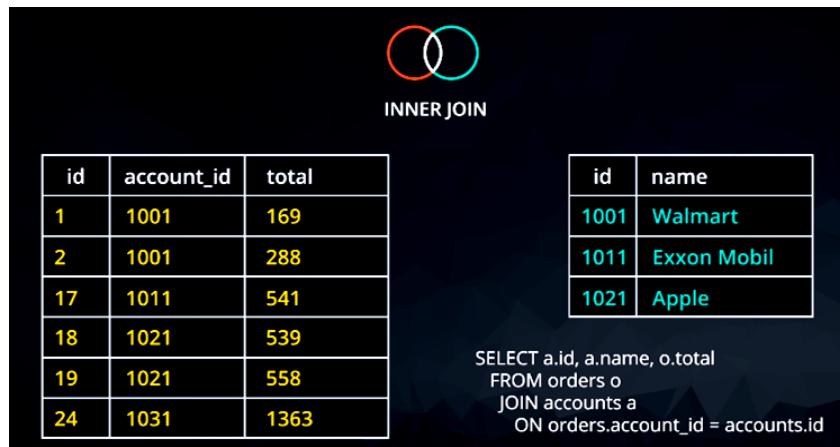
INNER JOIN RESULT

id	name	total
1001	Walmart	169
1001	Walmart	288
1011	Exxon Mobil	541
1021	Apple	539
1021	Apple	558

SELECT a.id, a.name, o.total
FROM orders o
JOIN accounts a
ON orders.account_id = accounts.id

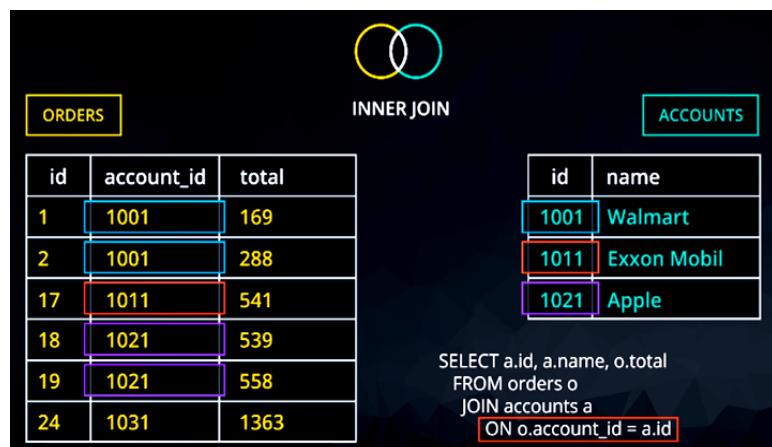
Notice we have only the three account IDs that were in both tables.

A common way to visualize different JOINS is with a venn diagram.



Each circle in the diagram represents table. The left circle includes all rows of data in the table in our FROM clause, which is the orders table. The right circle represents all rows of data in the table in our JOIN clause. In this case, that's the accounts table.

The overlapping middle section represent all rows for which the ON clause is true. An INNER JOIN, as shown here, will return only rows at the intersection of these two circles.



So what about this new account that hasn't placed any orders, yet ?

If we want to show that there are existing accounts that do not appear in the orders table, an **INNER JOIN** is not going to help us. In the upcoming sections, we'll work with other **OUTER JOINs** to pull that type information.

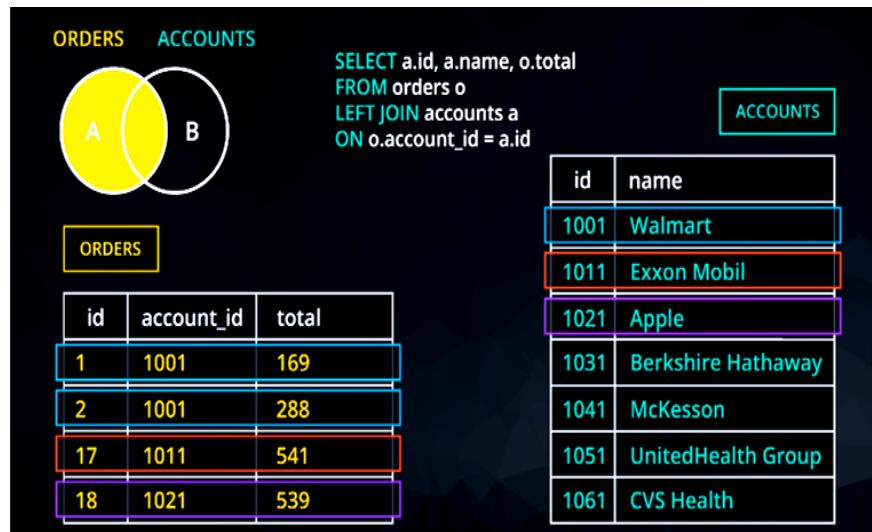
If we want to include data that doesn't exist in both tables, but only in one of the two tables we are using in our joint statement, **there are three types of joins we might use: a left join, a right join and a full outer join.**

Each of these joins will provide all the resulting rows of an INNER JOIN, but we may also gain some additional rows. So the result of an OUTER JOIN will always have at least as many rows as an INNER JOIN, with the same logic in the, on clause.

This lesson is going to focus primarily on LEFT and RIGHT JOINS. When we read our SELECT statements, the table listed in the FROM statement is considered the left table, while the table in the join considered the right table.

By simply writing the word , left. The result that we get back, will include of all the results that match with the right table, just like in an inner join. It will also return any results that are in the left table that did not match.

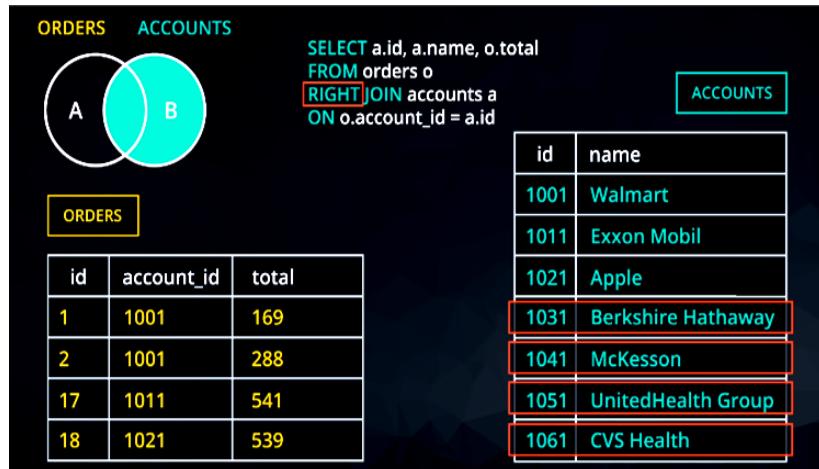
Using these subsets of the accounts and orders tables, every single row in the orders table here, can be matched to a row in the accounts table, with an ID in the accounts table the corresponds to an account ID in the orders table.



This result set exactly the same as if we had used an inner join.

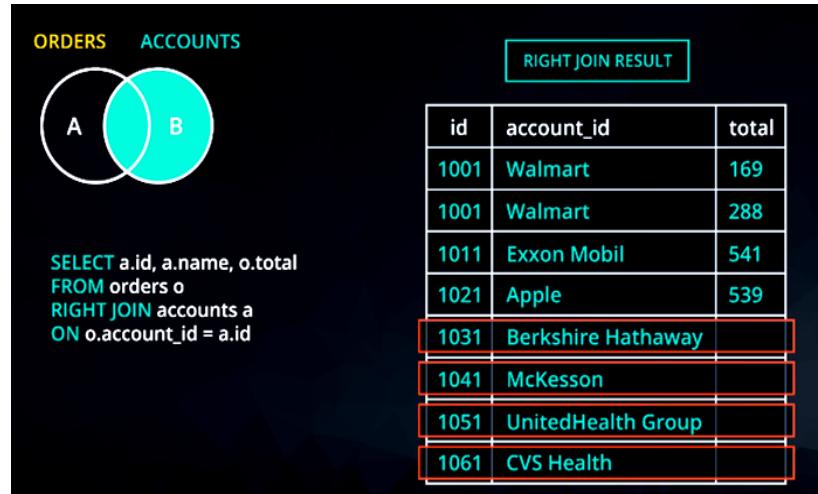


A right join would produce a different set of results, though. There are four accounts without any orders, which means four rows of data that do not satisfy the condition of this join.

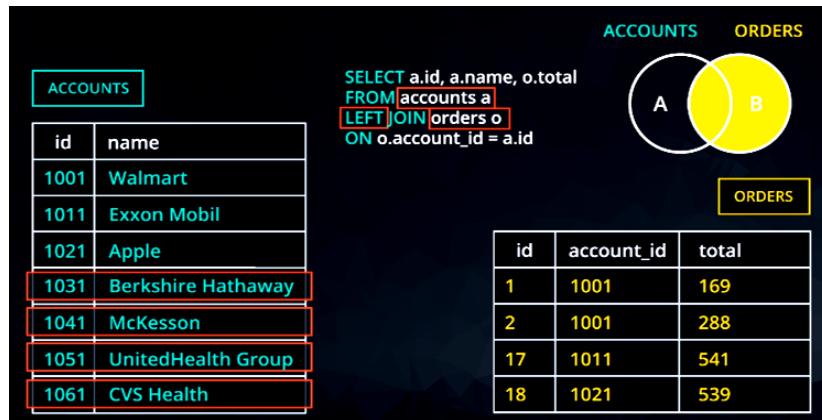


Since it's right join and not an inner join, these rows will be included at the bottom of the result set. They don't match with rows in the left table, so any columns from the left table will contain no data for these rows. Will discuss more about that this means soon enough.

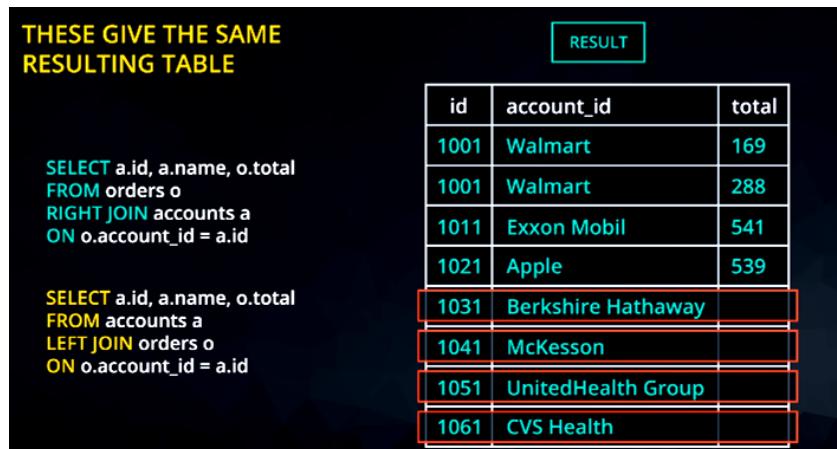
You can see that after we join, this final table has all the same rows as the original inner join table. But it has these extra rows from the accounts table at the bottom that were not matched with any orders.



Left and right joins are somewhat interchangeable. If we change this query so that the accounts table is in the FROM clause, and the orders table is in the join clause, the extra rows will be in the left table.



We can now run a left join, and the results will be exactly the same as the right join we did with the previous query.



Because left and right joins are effectively interchangeable, we use left joins through the rest of this course. The world is by and large standardized around this behavior, so you'll rarely see a right join in the wild. This makes it easier to read other people's queries, and quickly understand what they intended to accomplish.

Consistency is especially important if you're writing a query with multiple joins. We're not quite there yet, but we will be soon.

Other JOIN Notes

JOIN Check In

INNER JOINS

Notice **every** JOIN we have done up to this point has been an **INNER JOIN**. That is, we have always pulled rows only if they exist as a match across two tables.

Our new **JOINS** allow us to pull rows that might only exist in one of the two tables. This will introduce a new data type called **NULL**. This data type will be discussed in detail in the next lesson.

Quick Note

You might see the SQL syntax of

LEFT OUTER JOIN

OR

RIGHT OUTER JOIN

These are the exact same commands as the **LEFT JOIN** and **RIGHT JOIN** we learned about in the previous.

OUTER JOINS

The last type of join is a full outer join. This will return the inner join result set, as well as any unmatched rows from either of the two tables being joined.

Again this returns rows that **do not match** one another from the two tables. The use cases for a full outer join are **very rare**.

Similar to the above, you might see the language **FULL OUTER JOIN**, which is the same as **OUTER JOIN**.

★Quiz : LEFT and RIGHT JOIN

QUESTION 1 OF 4

Select all of the below that are true.

- A LEFT JOIN and RIGHT JOIN do the same thing if we change the tables that are in the FROM and JOIN statements.
- A LEFT JOIN will at least return all the rows that are in an INNER JOIN.
- JOIN and INNER JOIN are the same.
- A LEFT OUTER JOIN is the same as LEFT JOIN.

Country		State		
countryid	countryName	stateid	countryid	stateName
1	India	1	1	Maharashtra
2	Nepal	2	1	Punjab
3	United States	3	2	Kathmandu
4	Canada	4	3	California
5	Sri Lanka	5	3	Texas
6	Brazil	6	4	Alberta

Above are two small tables for you to test your knowledge of **JOINS**. You can click on the image to get a better view.

Country has 6 rows and 2 columns:

- **Country id** and **country Name**

State has 6 rows and 3 columns:

- **State id**, **country id**, and **state Name**

Use the above tables to determine the solution to the following questions.

QUESTION 2 OF 4

Match each statement to the item it describes.

Country.contryName

State.stateName

DESCRIPTION

ITEM

The primary key of the **Country** table.

Country.countryid

The primary key of the **State** table.

State.stateid

The foreign key that would be used in
JOINing the tables.

State.countryid

If you were to perform the following query:

```
SELECT c.countryid, c.countryName, s.stateName
FROM Country c
JOIN State s
ON c.countryid = s.countryid;
```

QUESTION 3 OF 4

Match the results of the query to the description.

8 5 4 7 1 12

DESCRIPTION	RESULT
The number of columns in resulting table.	3
The number of rows in the resulting table.	6
The number of times countryid 1 will show up in resulting table.	2
The number of times countryid 6 will show up in the resulting table.	0

Note :

That's right! Since this is a JOIN (INNER JOIN technically), we only get rows that show up in both tables. Therefore our resulting table will essentially look like the right table with the country Name pulled in as a column. Since 1, 2, 3, and 4 are country ids in both tables, this information will be pulled together. The country ids of 5 and 6 only show up in the Country table. Therefore, these will be dropped.

If you were to perform the following query:

```
SELECT c.countryid, c.countryName, s.stateName  
FROM Country c  
LEFT JOIN State s  
ON c.countryid = s.countryid;
```

QUESTION 4 OF 4

Match the results of the query to the description.

10 0 4 9 7 6 5

DESCRIPTION	RESULTS
The number of columns in resulting table.	3
The number of rows in the resulting table.	8
The number of times countryid 1 will show up in resulting table.	2
The number of times countryid 6 will show up in the resulting table.	1

NOTE :

That's right! We have a column for each of the identified elements in our **SELECT** statement. We will have all of the same rows as in a **JOIN** statement, but we also will obtain the additional two rows in the **Country** table that are not in the **State** table for **Sri Lanka** and **Brazil**.

JOINs and Filtering

A simple rule to remember this is that, when the database executes this query, it executes the join and everything in the **ON** clause first. Think of this as building the new result set. That result set is then filtered using the **WHERE** clause.

The fact that this example is a left join is important. Because inner joins only return the rows for which the two tables match, moving this filter to the **ON** clause of an inner join will produce the same result as keeping it in the **WHERE** clause.

The tricky thing about joins is that there are means to do other types of analysis. Combining data isn't an end unto itself. It's a tool that allows you to filter or aggregate with an expanded set of information. That means queries involving joins can get pretty complicated.

In order to get the exact results you're after, you need to be careful about exactly how you filter the data. As with joins, there are multiple points.

To illustrate this, let's take a look at all the orders that were brought in by one particular sales rep.

Example :

Imagine yourself as a sales manager at Parch & Posey. You'd like to make it easy for a sales rep to find her own deals rather than having to sift through all the orders and try and remember which were hers.

Remember, sales rep ID isn't in the orders table, so joins isn't necessary to get that information. The first way to filter this data is something we're already familiar with, the **WHERE** clause.

Let's take our combined orders and accounts tables and filter the results set to contain only the orders booked by sales rep 321500.

Input

HISTORY ▾ MENU ▾

SCHEMA

accounts

orders

region

sales_reps

web_events

```

1 SELECT o.* , a.*
2 FROM orders o
3 LEFT JOIN accounts a
4 ON o.account_id = a.id
5 WHERE a.sales_rep_id = 321500

```

Success!

EVALUATE

Output 134 results

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1001	1001	2015-10-06T17:31:14.000Z	123	22	24
1001	1001	2015-11-05T03:34:33.000Z	190	41	57
1001	1001	2015-12-04T04:21:55.000Z	85	47	0
1001	1001	2016-01-02T01:18:24.000Z	144	32	0

As you can see, the only rows in our result set are the rows that fit the criteria in the WHERE clause. The ones that have a sales rep ID equal to **321500**. But, what if we just want to mark all the orders that belong to rep **321500** and keep all the other orders in the result set as well ?

We can do that by filtering in the ON clause. Changing where to AND, we're moving this logical statement to become part of the ON clause.

Input

HISTORY ▾ MENU ▾

SCHEMA

accounts

orders

region

sales_reps

web_events

```

1 SELECT o.* , a.*
2 FROM orders o
3 LEFT JOIN accounts a
4 ON o.account_id = a.id
5 AND a.sales_rep_id = 321500

```

Success!

EVALUATE

Output 6912 results

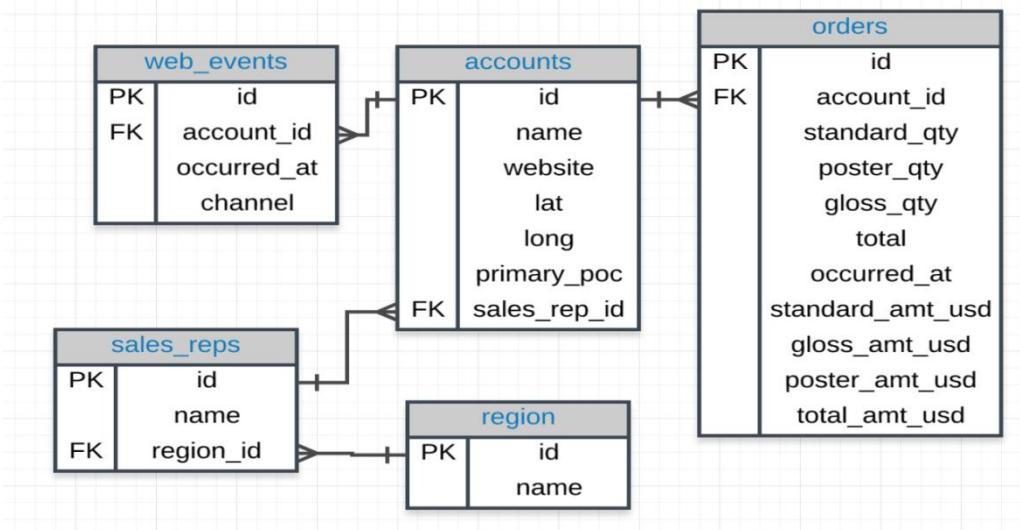
id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
1001	1001	2015-10-06T17:31:14.000Z	123	22	24
1001	1001	2015-11-05T03:34:33.000Z	190	41	57
1001	1001	2015-12-04T04:21:55.000Z	85	47	0
1001	1001	2016-01-02T01:18:24.000Z	144	32	0

This effectively pre-filters the right table to include only rows with sales rep ID **321500** before the join is executed. In another words, it's like a WHERE clause that applies before the join rather than after. You can almost think of this like joining orders to a different table. One that only includes a subset of the rows in the original accounts table.

Since this is a left join, we'll see result set that includes all rows from the orders table as well as any data in this new pre-filtered table that match with the account ID in the orders table.

As you can see, sales rep 321500 only had a few accounts. So, a lot of these orders remain unmatched. It might seem frivolous to leave so many orders in this result set without account names. But, this type of join will be incredibly useful once you move on to data aggregation in the next lesson.

Quiz: Last Check



Questions

- Provide a table that provides the **region** for each **sales_rep** along with their associated **accounts**. This time only for the **Midwest** region. Your final table should include three columns: the region **name**, the sales rep **name**, and the account **name**. Sort the accounts alphabetically (A-Z) according to account name.

Input

SCHEMA	▼
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1  SELECT r.name region, s.name rep, a.name account
2  FROM sales_reps s
3  JOIN region r
4  ON s.region_id = r.id
5  JOIN accounts a
6  ON a.sales_rep_id = s.id
7  WHERE r.name = 'Midwest'
8  ORDER BY a.name;

```

HISTORY ▾ MENU ▾

Success! EVALUATE

Output 48 results

region	rep	account
Midwest	Chau Rowles	Abbott Laboratories
Midwest	Julie Starr	AbbVie
Midwest	Cliff Meints	Aflac

2. Provide a table that provides the **region** for each **sales_rep** along with their associated **accounts**. This time only for accounts where the sales rep has a first name starting with **S** and in the **Midwest** region. Your final table should include three columns: the region **name**, the sales rep **name**, and the account **name**. Sort the accounts alphabetically (A-Z) according to account name.

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

HISTORY ▾ MENU ▾

```
1 SELECT r.name region, s.name rep, a.name account
2 FROM sales_reps s
3 JOIN region r
4 ON s.region_id = r.id
5 JOIN accounts a
6 ON a.sales_rep_id = s.id
7 WHERE r.name = 'Midwest' AND s.name LIKE 'S%'
8 ORDER BY a.name;
```

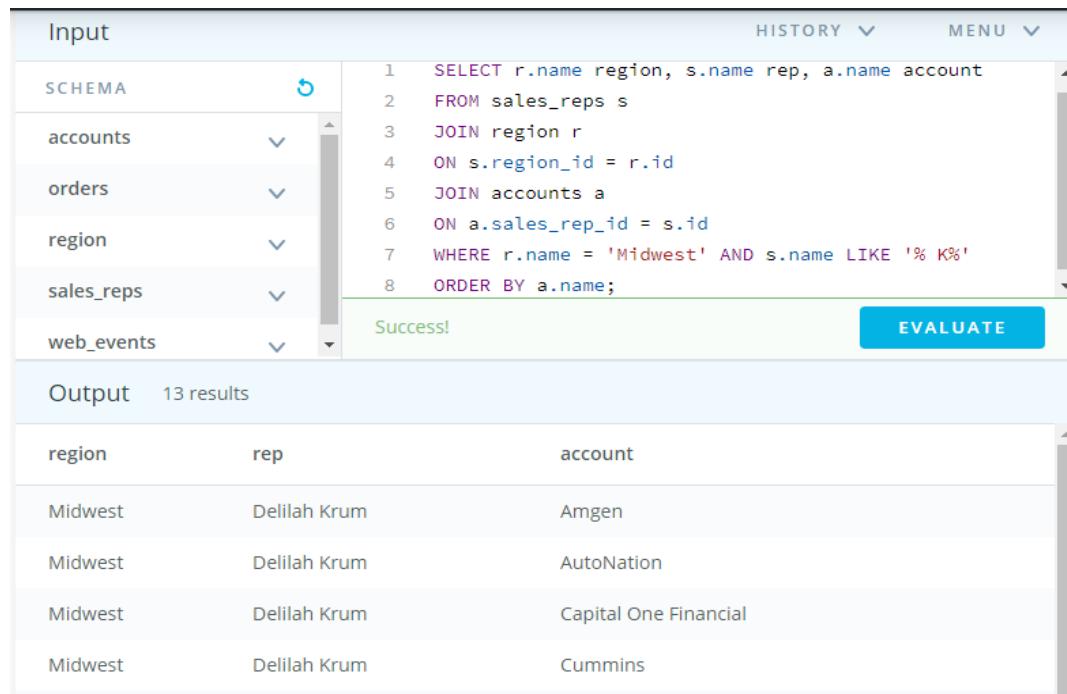
EVALUATE

Success!

Output 5 results

region	rep	account
Midwest	Sherlene Wetherington	Community Health Systems
Midwest	Sherlene Wetherington	Progressive
Midwest	Sherlene Wetherington	Rite Aid

3. Provide a table that provides the **region** for each **sales_rep** along with their associated **accounts**. This time only for accounts where the sales rep has a **last** name starting with **K** and in the **Midwest** region. Your final table should include three columns: the region **name**, the sales rep **name**, and the account **name**. Sort the accounts alphabetically (A-Z) according to account name.



The screenshot shows a database query interface with the following details:

- Input:** A schema browser on the left lists **accounts**, **orders**, **region**, **sales_reps**, and **web_events**.
- Query Editor:** The code area contains the following SQL query:


```

1  SELECT r.name region, s.name rep, a.name account
2  FROM sales_reps s
3  JOIN region r
4  ON s.region_id = r.id
5  JOIN accounts a
6  ON a.sales_rep_id = s.id
7  WHERE r.name = 'Midwest' AND s.name LIKE '% K%'
8  ORDER BY a.name;
      
```
- Status:** The message "Success!" is displayed below the query.
- EVALUATE:** A blue button is located to the right of the status message.
- Output:** A table titled "Output 13 results" shows the following data:

region	rep	account
Midwest	Delilah Krum	Amgen
Midwest	Delilah Krum	AutoNation
Midwest	Delilah Krum	Capital One Financial
Midwest	Delilah Krum	Cummins

4. Provide the **name** for each region for every **order**, as well as the account **name** and the **unit price** they paid ($\text{total_amt_usd}/\text{total}$) for the order. However, you should only provide the results if the **standard order quantity** exceeds **100**. Your final table should have 3 columns: **region name**, **account name**, and **unit price**. In order to avoid a division by zero error, adding .01 to the denominator here is helpful $\text{total_amt_usd}/(\text{total}+0.01)$.

The screenshot shows a database query interface with the following details:

- Input Schema:** The user has selected the following schemas: accounts, orders, region, sales_reps, and web_events.
- Query:**

```

1  SELECT r.name region, a.name account,
2    o.total_amt_usd/(o.total + 0.01) unit_price
3  FROM region r
4    JOIN sales_reps s
5      ON s.region_id = r.id
6    JOIN accounts a
7      ON a.sales_rep_id = s.id
8    JOIN orders o
9      ON o.account_id = a.id
  WHERE o.standard_qty > 100;
  
```
- Output:** The query was successful, returning 4509 results.
- Table Data:** The results are presented in a table with three columns: region, account, and unit_price. The data shows multiple rows for the Northeast region, all associated with the account "Walmart".

region	account	unit_price
Northeast	Walmart	5.7596000236672386
Northeast	Walmart	5.9651748203187389
Northeast	Walmart	5.4442361229475598
Northeast	Walmart	5.9601842312587116
Northeast	Walmart	6.1687185711808566
Northeast	Walmart	6.6289102252112738

5. Provide the **name** for each region for every **order**, as well as the account **name** and the **unit price** they paid ($\text{total_amt_usd}/\text{total}$) for the order. However, you should only provide the results if the **standard order quantity** exceeds **100** and the **poster order quantity** exceeds **50**. Your final table should have 3 columns: **region name**, **account name**, and **unit price**. Sort for the smallest **unit price** first. In order to avoid a division by zero error, adding .01 to the denominator here is helpful ($\text{total_amt_usd}/(\text{total}+0.01)$).

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales reps	▼
web_events	▼

```

1  SELECT r.name region, a.name account,
2      o.total_amt_usd/(o.total + 0.01) unit_price
3  FROM region r
4  JOIN sales_reps s
5  ON s.region_id = r.id
6  JOIN accounts a
7  ON a.sales_rep_id = s.id
8  JOIN orders o
9  ON o.account_id = a.id
10 WHERE o.standard_qty > 100 AND o.poster_qty > 50
11 ORDER BY unit_price;

```

Success! EVALUATE

Output 835 results

region	account	unit_price
Northeast	State Farm Insurance Cos.	5.1192822502542913
Southeast	DISH Network	5.2318158475403638
Northeast	Travelers Cos.	5.2351813313106532
Northeast	Best Buy	5.2604264379300265
West	Stanley Black & Decker	5.2663955600739988
Northeast	Citigroup	5.2730812103197040
Southeast	BlackRock	5.2782700457278773
Southeast	Nucor	5.2890939504788515

6. Provide the **name** for each region for every **order**, as well as the account **name** and the **unit price** they paid ($\text{total_amt_usd}/(\text{total} + 0.01)$) for the order. However, you should only provide the results if the **standard order quantity** exceeds **100** and the **poster order quantity** exceeds **50**. Your final table should have 3 columns: **region name**, **account name**, and **unit price**. Sort for the largest **unit price** first. In order to avoid a division by zero error, adding .01 to the denominator here is helpful ($\text{total_amt_usd}/(\text{total}+0.01)$)

Input

```

HISTORY ▾ MENU ▾
SCHEMA
accounts
orders
region
sales_reps
web_events
1  SELECT r.name region, a.name account,
2    o.total_amt_usd/(o.total + 0.01) unit_price
3  FROM region r
4  JOIN sales_reps s
5  ON s.region_id = r.id
6  JOIN accounts a
7  ON a.sales_rep_id = s.id
8  JOIN orders o
9  ON o.account_id = a.id
10 WHERE o.standard_qty > 100 AND o.poster_qty > 50
11 ORDER BY unit_price DESC;
Success!
EVALUATE

```

Output 835 results

region	account	unit_price
Northeast	IBM	8.0899060822781456
West	Mosaic	8.0663292103581285
West	Pacific Life	8.0630226525147913
Northeast	CHS	8.0188493267801133
West	Fidelity National Financial	7.9928024668468328
Midwest	Paccar	7.9869617587185754
Southeast	PNC Financial Services Group	7.8951386043342342
Northeast	Costco	7.7893519002932727

^MENU

7. What are the different **channels** used by **account id 1001**? Your final table should have only 2 columns: **account name** and the different **channels**. You can try **SELECT DISTINCT** to narrow down the results to only the unique values.

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1  SELECT DISTINCT a.name, w.channel
2  FROM accounts a
3  RIGHT JOIN web_events w
4  ON a.id = w.account_id
5  WHERE a.id = '1001';
```

Success!

EVALUATE

Output 6 results

name	channel
Walmart	adwords
Walmart	banner
Walmart	direct
Walmart	facebook
Walmart	organic

8. Find all the orders that occurred in **2015**. Your final table should have 4 columns: **occurred_at**, **account name**, **order total**, and **order total_amt_usd**.

Input

```
SCHEMA          1  SELECT o.occurred_at, a.name, o.total, o.total_amt_usd
accounts        2  FROM accounts a
orders          3  JOIN orders o
region          4  ON o.account_id = a.id
sales_reps      5  WHERE o.occurred_at BETWEEN '01-01-2015' AND '01-01-
web_events      6  ORDER BY o.occurred_at DESC;
```

Success!

EVALUATE

Output 1725 results

occurred_at	name	total	total_amt_usd
2015-12-31T23:21:15.000Z	Thermo Fisher Scientific	61	446.97
2015-12-31T23:15:35.000Z	Thermo Fisher Scientific	635	3246.90
2015-12-31T20:44:28.000Z	Coca-Cola	528	2693.54
2015-12-31T15:12:41.000Z	Computer Sciences	164	875.25
2015-12-31T15:11:15.000Z	Cameron International	513	2626.82

Recap & Looking Ahead

Primary and Foreign Keys

You learned a key element for **JOIN**ing tables in a database has to do with primary and foreign keys:

- **primary keys** - are unique for every row in a table. These are generally the first column in our database (like you saw with the **id** column for every table in the Parch & Posey database).
- **foreign keys** - are the **primary key** appearing in another table, which allows the rows to be non-unique.

Choosing the set up of data in our database is very important, but not usually the job of a data analyst. This process is known as **Database Normalization**.

JOINS

In this lesson, you learned how to combine data from multiple tables using **JOINS**. The three **JOIN** statements you are most likely to use are:

1. **JOIN** - an **INNER JOIN** that only pulls data that exists in both tables.
2. **LEFT JOIN** - a way to pull all of the rows from the table in the **FROM** even if they do not exist in the **JOIN** statement.
3. **RIGHT JOIN** - a way to pull all of the rows from the table in the **JOIN** even if they do not exist in the **FROM** statement.

There are a few more advanced **JOINS** that we did not cover here, and they are used in very specific use cases. [UNION](#) and [UNION ALL](#), [CROSS JOIN](#), and the tricky [SELF JOIN](#). These are more advanced than this course will cover, but it is useful to be aware that they exist, as they are useful in special cases.

Alias

You learned that you can alias tables and columns using **AS** or not using it. This allows you to be more efficient in the number of characters you need to write, while at the same time you can assure that your column headings are informative of the data in your table.

Looking Ahead

The next lesson is aimed at **aggregating** data. You have already learned a ton, but **SQL** might still feel a bit disconnected from **statistics** and using **Excel** like platforms. Aggregations will allow you to write **SQL** code that will allow for more complex queries, which assist in answering questions like:

- Which **channel** generated more revenue?
- Which **account** had an order with the most items?
- Which **sales_rep** had the most orders? or least orders? How many orders did they have?

Introduction to Aggregation

In the following concepts you will be learning in detail about each of the aggregate functions mentioned as well as some additional aggregate functions that are used in SQL all the time.

So far in this course, we've used SQL to join tables using left, right and inner joins, and filter the results using Where and On clauses.

Sometimes, it's useful to view row-level data. For example, we might want to look at all the individual orders for a given sales person to see how they're performing in a given month. But other times, row-level data is overwhelming, and ultimately, less valuable than aggregated data. Counting up all the orders in each region in each month, for example, will be pretty inconvenient when looking at row-level data.

One rep's order volume is low enough to look at individual rows, but the entire region's volume is way too high. Instead, a single number of orders for each month will be much better.

Fortunately, databases are great at aggregating data. The SQL syntax is easy to learn. If you're familiar with Excel, SQL commands are pretty similar.

- ✚ Count, counts how many rows are in a particular column.
- ✚ Sum, adds together all the values in a particular column.
- ✚ Min and Max, return the lowest and highest values in a particular column.
- ✚ Average, calculates the average of all the values in a particular column.

These functions operate down columns, not across rows. So you can do things like sum up all the quantities of paper over delivered. Instead of getting results set many thousands of rows long, you just get one line with the answer.

In practice, you'll find yourself using row-level output for the early exploratory work when searching your database to better understand the data.

As you get a sense for what the data looks like and begin to look for answer to your questions, aggregates become more helpful.

Introduction to NULLs

NULLs are a data type that specifies where no data exists in SQL. They are often ignored in our aggregation functions, which you will get a first look at in the next concept using **COUNT**.

Before we dive into aggregation, we need to cover one more concept, NULLs.

NULL means no data. It's different from a zero or a space. And there's a good reason. From a business perspective, for Parch & Posey a zero means that no paper was sold, which could mean that a sale was attempted but not made.

A null could mean no sale was even attempted, and that's a pretty meaningful difference.

Those could be the result of any number of things, they can be part of the data design itself or they can simply be the result of poorly looked data.

Regardless, it's important to understand what nulls mean and to know how to work with them.

NULLs and Aggregation

Notice that **NULLs** are different than a zero - they are cells where data does not exist.

When identifying **NULLs** in a **WHERE** clause, we write **IS NULL** or **IS NOT NULL**. We don't use **=**, because **NULL** isn't considered a value in SQL. Rather, it is a property of the data.

NULLs - Expert Tip

There are two common ways in which you are likely to encounter **NULLs**:

- **NULLs** frequently occur when performing a **LEFT** or **RIGHT JOIN**. You saw in the last lesson - when some rows in the left table of a left join are not matched with rows in the right table, those rows will contain some **NULL** values in the result set.
- **NULLs** can also occur from simply missing data in our database.

Example:

To get a feel for how nulls work, let's pull a set of records from the accounts table with IDs that fall between 1500 and 1600.

The screenshot shows a SQL query interface with the following details:

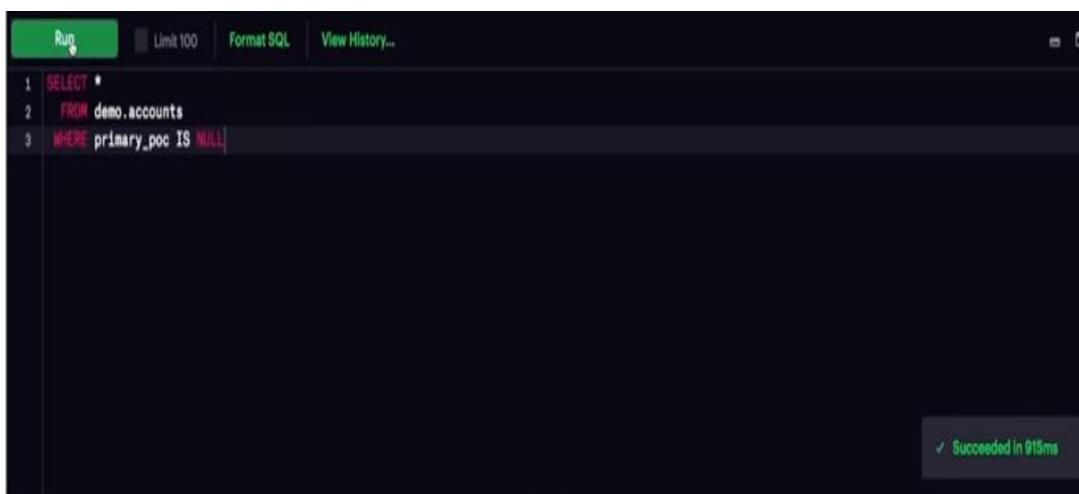
- Run** button (green)
- Limit 100** button
- Format SQL** button
- View History...** button
- SQL Editor:**

```
1 SELECT *
2   FROM demo.accounts
3 WHERE id > 1500 AND id < 1600;
```
- Status Bar:** ✓ Succeeded in 2s
- Results:** 10 rows returned
- Export, Copy, Chart, Pivot** buttons
- Table Headers:** id, name, website, lat, long, primary_poc, sales_rep_id
- Table Data:**

	id	name	website	lat	long	primary_poc	sales_rep_id
1	1501	Intel	www.intel.com	41.03153857	-74.66846407		321580
2	1511	Humana	www.humana.com	41.43233665	-77.00496342	Bettye Close	321590
3	1521	Disney	www.disney.com	41.87879976	-74.81102607	Timika Mistretta	321600
4	1531	Cisco Systems	www.cisco.com	41.20101093	-76.53824668	Deadra Waggener	321610

We can see that the primary_poc, which stands for point of contact, is blank for the Intel account. This could simply be an error in the data. Maybe a point of contact that was accidentally deleted at some point, or it could be that Parch & Posey's point of contact left the company and they don't yet have a new point of contact for that account. Either way, there is no data in this particular cell. This cell is null.

Imagine yourself as a sales manager at Parch & Posey, you may want to know all of the accounts for which the primary_poc is null. If you don't have a point of contact, chances are you're not going to be able to keep that customer for much longer. In order to find all these accounts, we'll have to use some special syntax in our WHERE clause.



The screenshot shows a SQL query editor with the following code:

```
Run Limit 100 Format SQL View History...
1 SELECT *
2 FROM demo.accounts
3 WHERE primary_poc IS NULL
```

A green success message at the bottom right says "✓ Succeeded in 015ms".

	id	name	website	lat	long	primary_poc	sales_rep_id
1	1501	Intel	www.intel.com	41.03153857	-74.66846407		321580
2	1671	Delta Air Lines	www.delta.com	40.75860903	-73.99067048		321510
3	1951	Twenty-First Century Fox	www.21cf.com	42.35467661	-71.05476697		321560
4	2131	USAA	www.usaa.com	41.87745439	-87.62793161		321780

Turns out, there are quite a few accounts without points of contact. So, the special thing about nulls that you need to write is NULL instead of EQUALS NULL.

A screenshot of a SQL query editor interface. The top bar has buttons for 'Run' (highlighted in green), 'Limit 100', 'Format SQL', and 'View History...'. The main area contains the following SQL code:

```
1 SELECT *
2 FROM demo.accounts
3 WHERE primary_poc = NULL
```

The status bar at the bottom right shows a green checkmark and the text 'Succeeded in 935ms'.

The result pane below displays the message 'No rows returned'.

The reason this doesn't work is that `NULL` is not a value, it's a property of the data. This is different from zero or a space, both of which are values.

If you want to find the inverse of our previous result set, you can use this syntax, `is not NULL`.

A screenshot of a SQL query editor interface. The top bar has buttons for 'Run' (highlighted in green), 'Limit 100', 'Format SQL', and 'View History...'. The main area contains the following SQL code:

```
1 SELECT *
2 FROM demo.accounts
3 WHERE primary_poc IS NOT NULL
```

The status bar at the bottom right shows a green checkmark and the text 'Succeeded in 1s'.

The result pane below shows a table with the following data:

	id	name	website	lat	long	primary_poc	sales_rep_id
1	1001	Walmart	www.walmart.com	40.23849561	-75.10329704	Tamara Tuma	321500
2	1011	Exxon Mobil	www.exxonmobil.com	41.1691563	-73.84937379	Sung Shields	321510
3	1021	Apple	www.apple.com	42.29049481	-76.08400942	Jodee Lupo	321520
4	1031	Berkshire Hathaway	www.berkshirehathaway.com	40.94902131	-75.76389759	Serafina Banda	321530

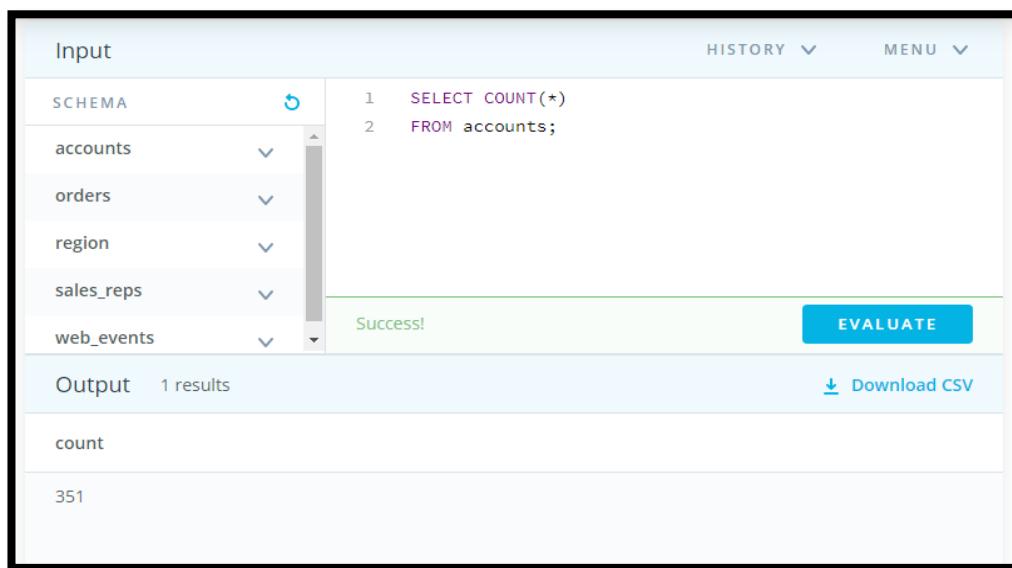
The total count '345 rows returned' is displayed at the bottom right of the result pane.

As you can see, this results all rows for which there are values in the `primary_poc` column.

First Aggregation - COUNT

COUNT the Number of Rows in a Table

Try your hand at finding the number of rows in each table. Here is an example of finding all the rows in the **accounts** table.

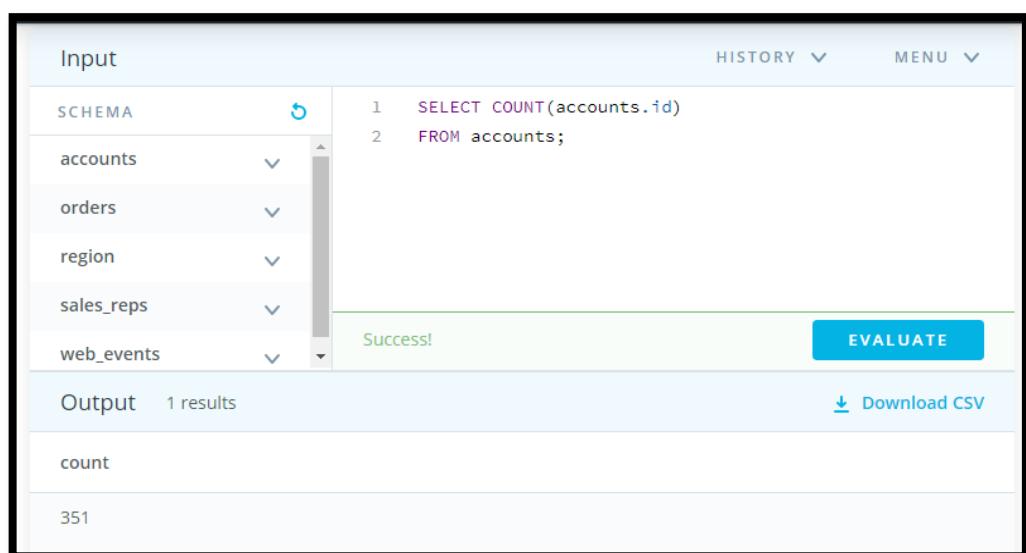


The screenshot shows a database query interface. In the 'Input' section, the schema dropdown is set to 'accounts'. The query entered is:

```
1 SELECT COUNT(*)
2 FROM accounts;
```

The output section shows 1 result: 'count' with the value '351'. A 'Success!' message is displayed next to the results. There is also a blue 'EVALUATE' button.

But we could have just as easily chosen a column to drop into the aggregation function:



The screenshot shows a database query interface. In the 'Input' section, the schema dropdown is set to 'accounts'. The query entered is:

```
1 SELECT COUNT(accounts.id)
2 FROM accounts;
```

The output section shows 1 result: 'count' with the value '351'. A 'Success!' message is displayed next to the results. There is also a blue 'EVALUATE' button.

These two statements are equivalent, but this isn't always the case.

Example :

Here's a list of orders in the last month, assuming it's currently January 2017. I'm using mode to run these queries, which tells me that there are 463 rows in my result set.

The screenshot shows a database interface with the following details:

Input:

- SCHEMA dropdown menu with options: accounts, orders, region, sales_reps, web_events.
- Query text area:

```
1 SELECT *
2 FROM orders o
3 WHERE occurred_at >= '2016-12-01' AND occurred_at <=
  '2017-01-01'
```
- Status message: Success!
- EVALUATE button.

Output:

- 463 results
- Download CSV link.
- Table view showing columns: id, account_id, occurred_at, standard_qty, gloss_qty, poster_qty.
- Sample data rows:

id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty
16	1001	2016-12-24T05:53:13.000Z	123	127	19
17	1011	2016-12-21T10:59:34.000Z	527	14	0
24	1031	2016-12-25T03:54:27.000Z	1148	0	215
27	1041	2016-12-12T07:50:01.000Z	157	34	21

Still, you'll need to learn to count in SQL, because you can use it count in a much more fine-grained way than simply counting all the rows in a result set.

We can do this by modifying our SELECT statement to use a COUNT function.

The screenshot shows a database interface with the following details:

Input:

- SCHEMA dropdown menu with options: accounts, orders, region, sales_reps, web_events.
- Query text area:

```
1 SELECT COUNT(*)
2 FROM orders o
3 WHERE occurred_at >= '2016-12-01' AND occurred_at <=
  '2017-01-01'
```
- Status message: Success!
- EVALUATE button.

Output:

- 1 results
- Download CSV link.
- Table view showing column: count.
- Value: 463

You can see that this returns a single row of output in a single column, and the value is exactly the same as the row COUNT we saw before.

Let's give that column an alias, so that it's a little easier to read.

The screenshot shows a SQL query editor interface. On the left, there is a sidebar titled "Input" with a "SCHEMA" dropdown containing "accounts", "orders", "region", "sales_reps", and "web_events". The main area displays a SQL query:

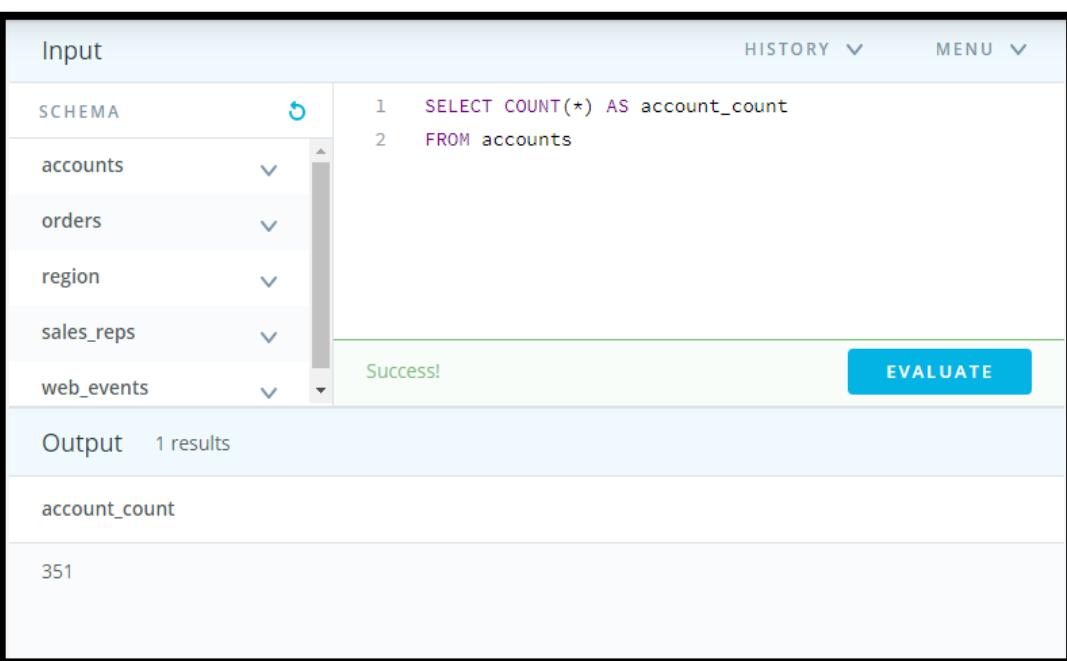
```
1 SELECT COUNT(*)
2 FROM orders o
3 WHERE occurred_at >= '2016-12-01' AND occurred_at <=
  '2017-01-01'
```

Below the query, a green "Success!" message is displayed next to a blue "EVALUATE" button. Under the "Output" section, it says "1 results" and shows a single row with the column "count" containing the value "463". There is also a "Download CSV" link.

The COUNT function is returning a count of all the rows that contain some non-null data. It's very unusual to have a row that is entirely null. So, the result produced by a **COUNT(*)** is typically equal to the number of rows in the table.

COUNT & NULLs

Notice that **COUNT** does not consider rows that have **NULL** values. Therefore, this can be useful for quickly identifying which rows have missing data. You will learn **GROUP BY** in an upcoming concept, and then each of these aggregators will become much more useful. The COUNT function can also be used to count the number of non-null records in an individual column. To illustrate this, let's take a look at the accounts table.



The screenshot shows a database interface with the following details:

- Input Schema:** accounts
- Query:**

```
1 SELECT COUNT(*) AS account_count
2 FROM accounts
```
- Status:** Success!
- Output:** 1 results
- Result:** account_count
351

As you can see, there are 354 rows in the accounts table. Now, substituting ID for the star, we can see how many non-nulls records there are in the ID column.

Let's give the column a more sensible name while we are at it.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT COUNT(id) AS account_id_count
2 FROM accounts
```

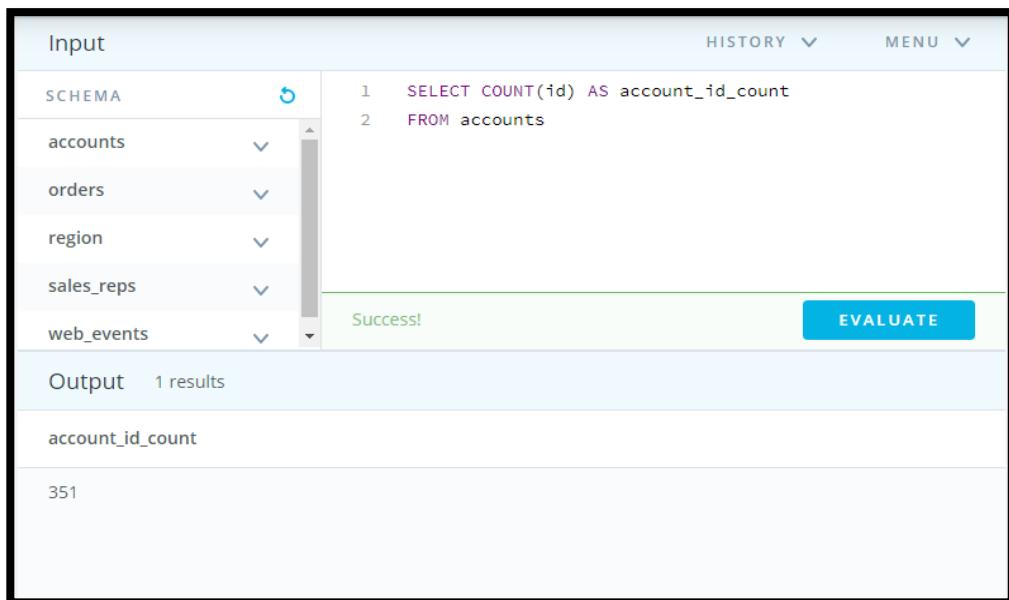
Success!

EVALUATE

Output 1 results

account_id_count

351



Since there are no null values in the ID column, it return the same result as count star.

Now, let's try this with a column that we know contains some null values, primary_poc.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT COUNT(primary_poc) AS account_primary_poc_count
2 FROM accounts
```

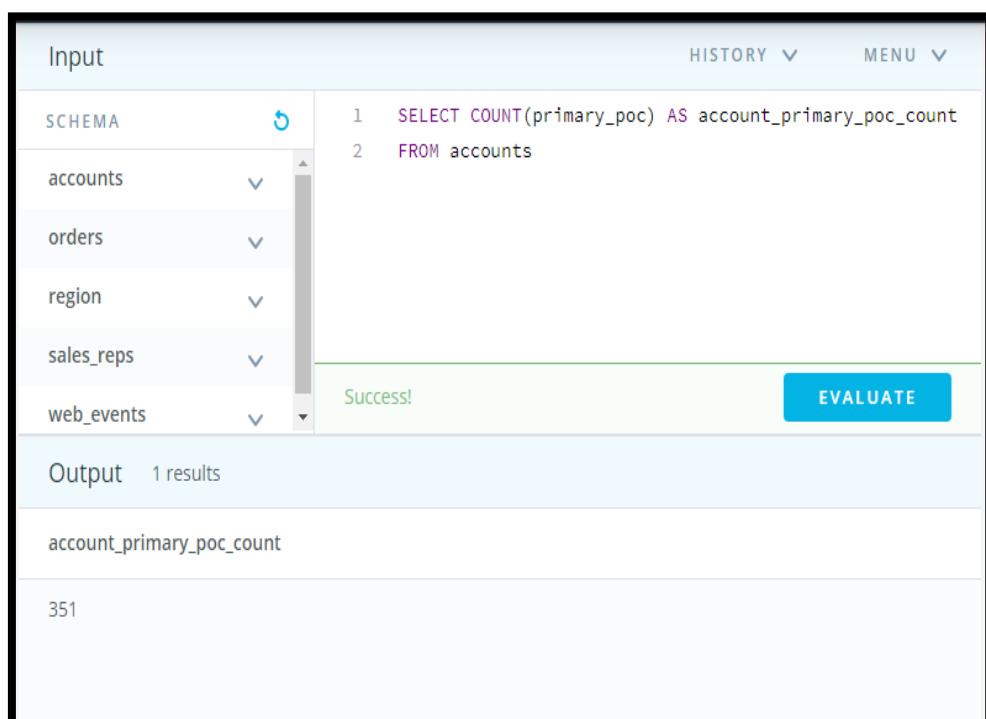
Success!

EVALUATE

Output 1 results

account_primary_poc_count

351



This time, we get a result than is nine lower than our previous result.

Let's run a quick query to verify that there are nine null values in the primary_poc column.

The screenshot shows a SQL query interface with the following details:

- Run** button is highlighted.
- SQL code:

```
1 SELECT *
2 FROM demo.accounts
3 WHERE primary_poc IS NULL
```
- Execution status: **Succeeded in 93ms**.
- Table results:

	id	name	website	lat	long	primary_poc	sales_rep_id
1	1501	Intel	www.intel.com	41.03153857	-74.66846407		321580
2	1671	Delta Air Lines	www.delta.com	40.75860903	-73.99067048		321510
3	1951	Twenty-First Century Fox	www.2icf.com	42.35467661	-71.05476697		321560
4	2131	USAA	www.usaa.com	41.87745439	-87.62793161		321780
- Bottom navigation: Export, Copy, Chart, Pivot.
- Total rows returned: 9.

And here they are, the nine rows with null values. One thing you may have noticed is that count can be used in columns with non numerical values. This is not true of all aggregation function, but it make sense.

The count function is just looking for non-null data and text is not null. Some of the functions we're about to learn like sum and average are impossible to apply to next because, well, how do you take the average of a bunch of account names.

SUM

Unlike **COUNT**, you can only use **SUM** on numeric columns. However, **SUM** will ignore **NULL** values, as do the other aggregation functions you will see in the upcoming lessons.

Aggregation Reminder

An important thing to remember: **aggregators only aggregate vertically - the values of a column**. If you want to perform a calculation across rows.

We saw this in the first lesson if you need a refresher, but the quiz in the next concept should assure you still remember how to aggregate across rows.

Example :

Imagine yourself as an operations manager at Parch & Posey. You're trying to do some inventory planning, and you want to know how much of each paper type of produce.

A good place to start might be to total up all sales of each paper type and compare them to one another. We'll do this using **SUM**. It works similar to **COUNT**, except that you'll want to specify column names rather than using star.

The screenshot shows a database interface with the following details:

- Input:** A sidebar on the left lists tables: SCHEMA, accounts, orders, region, sales_reps, and web_events.
- Query Editor:** The main area contains the following SQL code:

```
1 SELECT SUM(standard_qty) AS standard, SUM(gloss_qty) AS
2 gloss, SUM(poster_qty) AS poster
3 FROM orders
```
- Status:** A message "Success!" is displayed below the query.
- EVALUATE:** A blue button is located to the right of the status message.
- Output:** A table titled "Output" shows the results of the query:

standard	gloss	poster
1938346	1013773	723646

It looks like standard is more popular than both of the non-standard paper types combined. Unlike COUNT, you can only use the SUM function on columns containing numerical values. You don't need to worry too much about the presence of nulls. The SUM function will just start nulls as zero.

Quiz: SUM

Aggregation Questions

Use the **SQL** environment below to find the solution for each of the following questions. If you get stuck or want to check your answers, you can find the answers at the top of the next concept.

1. Find the total amount of **poster_qty** paper ordered in the **orders** table.

The screenshot shows a SQL environment interface. On the left, there is a sidebar titled "Input" with a "SCHEMA" dropdown containing options: accounts, orders, region, sales_reps, and web_events. The "orders" option is currently selected. On the right, the main area has "HISTORY" and "MENU" buttons at the top. Below them, a code editor displays the following SQL query:

```
1  SELECT SUM(poster_qty) AS total_poster_sales
2  FROM orders;
```

The status bar at the bottom of the code editor says "Success!". To the right of the status bar is a blue "EVALUATE" button. Below the code editor, there is an "Output" section with "1 results". The output table contains one row with the column "total_poster_sales" and the value "723646". At the bottom right of the output section is a link "Download CSV".

2. Find the total amount of **standard_qty** paper ordered in the **orders** table.

Input

HISTORY ▾ MENU ▾

SCHEMA

- accounts
- orders
- region
- sales_reps
- web_events

```
1 SELECT SUM(standard_qty) AS total_standard_sales
2 FROM orders;
```

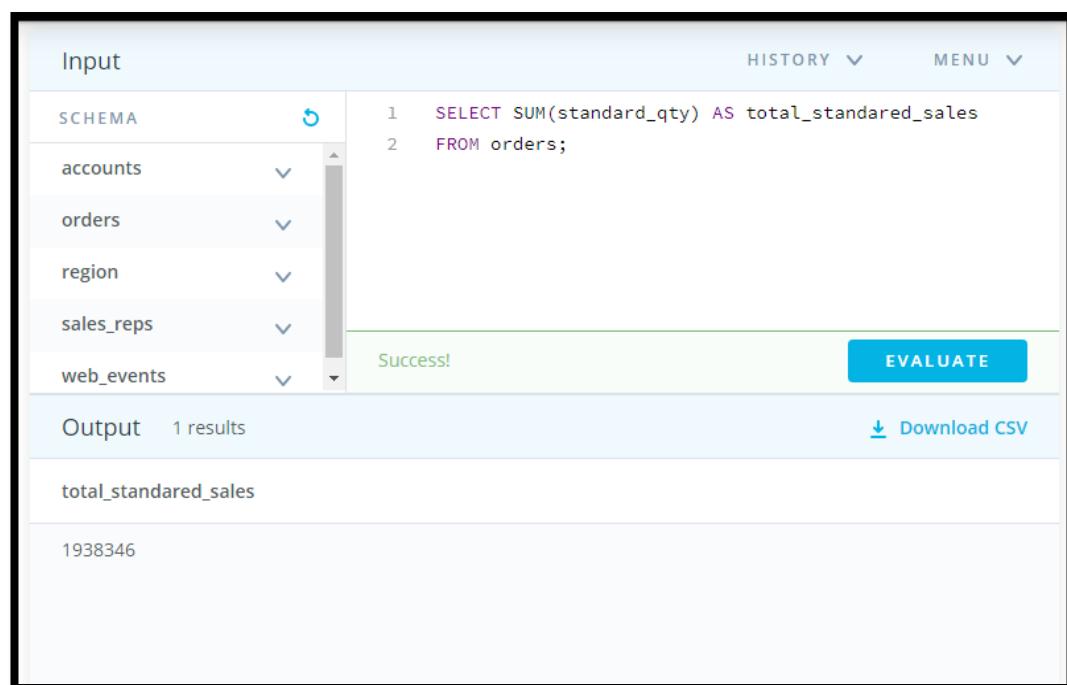
Success!

EVALUATE

Output 1 results

total_standard_sales

1938346



3. Find the total dollar amount of sales using the **total_amt_usd** in the **orders** table.

Input

HISTORY ▾ MENU ▾

SCHEMA

- accounts
- orders
- region
- sales_reps
- web_events

```
1 SELECT SUM(total_amt_usd) AS total_dollar_sales
2 FROM orders;
```

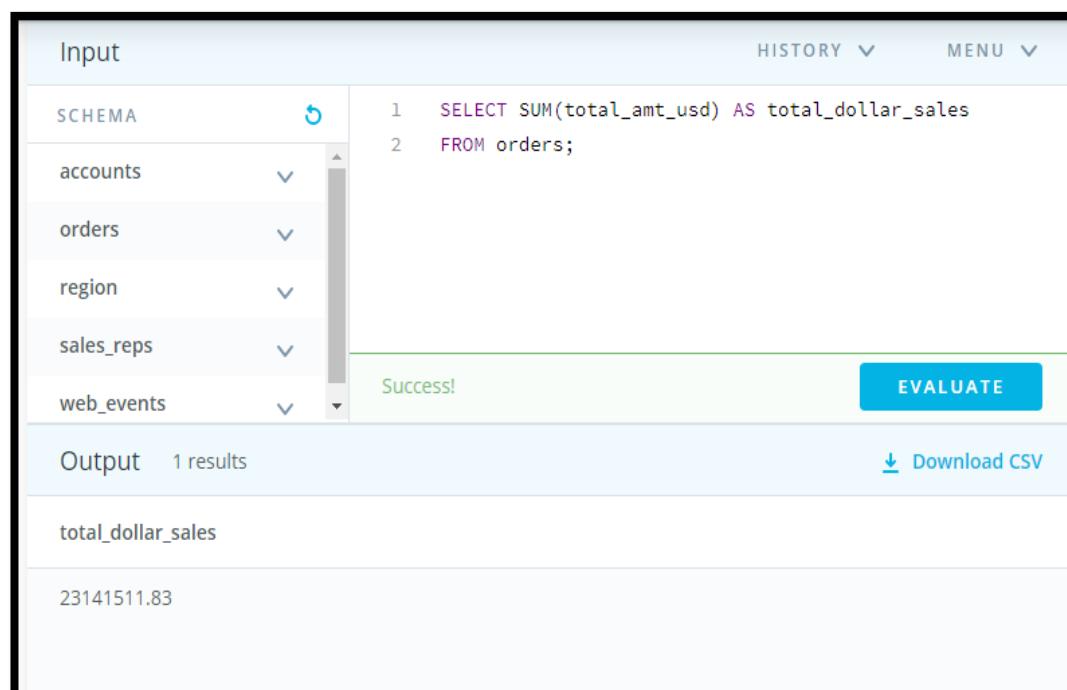
Success!

EVALUATE

Output 1 results

total_dollar_sales

23141511.83



4. Find the total amount spent on **standard_amt_usd** and **gloss_amt_usd** paper for each order in the orders table. This should give a dollar amount for each order in the table.

The screenshot shows a database query interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```
1 SELECT standard_amt_usd + gloss_amt_usd AS
  2   total_standard_gloss
    FROM orders;
```
- Status:** Success!
- Output:** 6912 results
- Results:** A list of values under the heading "total_standard_gloss":
 - 778.55
 - 1255.19
 - 776.18
 - 958.24
 - 756.13
 - 693.73
- Buttons:** EVALUATE, Download CSV

5. Find the **standard_amt_usd** per unit of **standard_qty** paper. Your solution should use both an aggregation and a mathematical operator.

The screenshot shows a database query interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```
1 SELECT SUM(standard_amt_usd)/SUM(standard_qty) AS
  2   standard_price_per_unit
    FROM orders;
```
- Status:** Success!
- Output:** 1 results
- Results:** A list of values under the heading "standard_price_per_unit":
 - 4.990000000000000
- Buttons:** EVALUATE, Download CSV

MIN & MAX



The screenshot shows a SQL editor interface with a green 'Run' button at the top left. To its right are 'Limit 100', 'Format SQL', and 'View History...' buttons. The main area contains the following SQL code:

```
1 SELECT MIN(standard_qty) AS standard_min,
2       MIN(gloss_qty) AS gloss_min,
3       MIN(poster_qty) AS poster_min,
4       MAX(standard_qty) AS standard_max,
5       MAX(gloss_qty) AS gloss_max,
6       MAX(poster_qty) AS poster_max
7 FROM demo.orders
```

Notice that here we were simultaneously obtaining the **MIN** and **MAX** number of orders of each paper type. However, you could run each individually.

Notice that **MIN** and **MAX** are aggregators that again ignore **NULL** values.

Expert Tip

Functionally, **MIN** and **MAX** are similar to **COUNT** in that they can be used on non-numerical columns. Depending on the column type, **MIN** will return the lowest number, earliest date, or non-numerical value as early in the alphabet as possible. As you might suspect, **MAX** does the opposite—it returns the highest number, the latest date, or the non-numerical value closest alphabetically to “Z.”

The syntax for min & max is similar to SUM & COUNT. It shouldn't be surprising that the minimum for each paper type is zero.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT MIN(standard_qty) AS standard_min,  
MIN(gloss_qty) AS gloss_min, MIN(poster_qty) AS  
poster_min, MAX(standard_qty) AS standard_max,  
MAX(gloss_qty) AS gloss_max, MAX(poster_qty) AS  
poster_max  
2 FROM orders
```

Success! EVALUATE

Output 1 results

standard_min	gloss_min	poster_min	standard_max	gloss_max	poster_max
0	0	0	22591	14281	28262

Some customers only order one or two types of paper. What is surprising is that the largest single order is for poster paper, despite the fact that it's the least popular overall.

There are important implications here. Even though it's not the most popular item, Parch & Posey may want to produce enough poster paper to be able to fulfill pretty big orders at any given time.

AVG

Similar to other software **AVG** returns the mean of the data - that is the sum of all of the values in the column divided by the number of values in a column. This aggregate function again ignores the **NULL** values in both the numerator and the denominator.

If you want to count **NULLs** as zero, you will need to use **SUM** and **COUNT**. However, this is probably not a good idea if the **NULL** values truly just represent unknown values for a cell.

MEDIAN - Expert Tip

One quick note that a median might be a more appropriate measure of center for this data, but finding the median happens to be a pretty difficult thing to get using SQL alone — so difficult that finding a median is occasionally asked as an interview question.

So now we know which paper types are most are most popular and we have a sense of largest order size we might need to fulfill at any given time. But what's the average order size?

What can we expect to see on a regular bases ?

We'll use the average function which is typed as **AVG** and has a similar syntax to all of the other aggregation functions.

The screenshot shows a SQL query editor interface. On the left, there is a sidebar titled "Input" with a "SCHEMA" dropdown set to "accounts". Below it are dropdowns for "orders", "region", "sales_reps", and "web_events". The main area contains a code editor with the following SQL query:

```
1 SELECT AVG(standard_qty) AS standard_avg,
2          AVG(gloss_qty) AS gloss_avg, AVG(poster_qty) AS
3          poster_avg
4 FROM orders
```

Below the code, a green "Success!" message is displayed next to a blue "EVALUATE" button. The "Output" section shows one result row:

standard_avg	gloss_avg	poster_avg
280.4320023148148148	146.6685474537037037	104.6941550925925926

So, it's look like that large poster paper order most have been a major outlier because the average poster order is only about a third the size of the average standard order.

When you're using the average function, keep in mind that it can only be used on numerical columns. Also, ignores nulls completely, meaning that rows with null values are not counted in the numerator or the denominator when calculating the average.

If you want to treat nulls as zero, you'll need to take a sum and divide it by the count rather than just using the average function.

Quiz: MIN, MAX, & AVG

Questions: MIN, MAX, & AVERAGE

Use the **SQL** environment below to assist with answering the following questions. Whether you get stuck or you just want to double check your solutions, my answers can be found at the top of the next concept.

1. When was the earliest order ever placed? You only need to return the date.

The screenshot shows a SQL environment interface. On the left, there's a sidebar titled "Input" with a "SCHEMA" dropdown containing "accounts", "orders", "region", "sales_reps", and "web_events". The "orders" schema is selected. On the right, the main area has "HISTORY" and "MENU" buttons at the top. Below them, a code editor shows two lines of SQL:

```
1 SELECT MIN(occurred_at)
2 FROM orders
```

A "Success!" message is displayed below the code. There's a blue "EVALUATE" button. Underneath, an "Output" section shows "1 results" and a "Download CSV" link. The output table contains one row with the value "min" and "2013-12-04T04:22:44.000Z".

2. Try performing the same query as in question 1 without using an aggregation function.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT occurred_at
2 FROM orders
3 ORDER BY occurred_at
4 LIMIT 1;
```

Success! EVALUATE

Output 1 results [Download CSV](#)

occurred_at
2013-12-04T04:22:44.000Z

3. When did the most recent (latest) **web_event** occur?

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT MAX (occurred_at)
2 FROM web_events
```

Success! EVALUATE

Output 1 results [Download CSV](#)

max
2017-01-01T23:51:09.000Z

4. Try to perform the result of the previous query without using an aggregation function.

The screenshot shows a database query interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events
- Query:**

```

1  SELECT occurred_at
2  FROM web_events
3  ORDER BY occurred_at desc
4  LIMIT 1;
    
```

- Status:** Success!
- EVALUATE** button
- Output:** 1 results
- Download CSV** link
- occurred_at** column data: 2017-01-01T23:51:09.000Z

5. Find the mean (**AVERAGE**) amount spent per order on each paper type, as well as the mean amount of each paper type purchased per order. Your final answer should have 6 values - one for each paper type for the average number of sales, as well as the average amount.

The screenshot shows a database query interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events
- Query:**

```

1  SELECT AVG(standard_qty) mean_standard, AVG(gloss_qty)
   mean_gloss, AVG(poster_qty) mean_poster,
   AVG(standard_amt_usd) mean_standard_usd,
   AVG(gloss_amt_usd) mean_gloss_usd, AVG(poster_amt_usd)
   mean_poster_usd
2  FROM orders;
    
```

- Status:** Success!
- EVALUATE** button
- Output:** 1 results
- Download CSV** link
- mean_standard**, **mean_gloss**, **mean_poster**, **mean_standard_usd** columns data:

280.4320023148148148	146.6685474537037037	104.6941550925925926	1399.35569155092592592
----------------------	----------------------	----------------------	------------------------

6. Via the video, you might be interested in how to calculate the MEDIAN. Though this is more advanced than what we have covered so far try finding - what is the MEDIAN **total_usd** spent on all **orders**?

The screenshot shows a database interface with the following details:

Input:

- SCHEMA dropdown menu with options: accounts, orders, region, sales_reps, web_events.
- SQL code entered in the main area:

```
1 SELECT *
2 FROM (SELECT total_amt_usd
3        FROM orders
4        ORDER BY total_amt_usd
5        LIMIT 3457) AS Table1
6 ORDER BY total_amt_usd DESC
7 LIMIT 2;
```
- A "Success!" message is displayed below the code.
- A blue "EVALUATE" button is located to the right of the message.

Output:

- 2 results
- total_amt_usd
- 2483.16
- 2482.55

At the bottom right, there is a "Download CSV" link.

GROUP BY

The key takeaways here:

- **GROUP BY** can be used to aggregate data within subsets of the data. For example, grouping for different accounts, different regions, or different sales representatives.
- Any column in the **SELECT** statement that is not within an aggregator must be in the **GROUP BY** clause.
- The **GROUP BY** always goes between **WHERE** and **ORDER BY**.
- **ORDER BY** works like **SORT** in spreadsheet software.

GROUP BY - Expert Tip

Before we dive deeper into aggregations using **GROUP BY** statements, it is worth noting that SQL evaluates the aggregations before the **LIMIT** clause. If you don't group by any columns, you'll get a 1-row result—no problem there. If you group by a column with enough unique values that it exceeds the **LIMIT** number, the aggregates will be calculated, and then some rows will simply be omitted from the results.

This is actually a nice way to do things because you know you're going to get the correct aggregates. If SQL cuts the table down to 100 rows, then performed the aggregations, your results would be substantially different. The above query's results exceed 100 rows, so it's a perfect example. In the next concept, use the SQL environment to try removing the **LIMIT** and running it again to see what changes.

So far, we've applied these aggregation functions across an entire table. You now know how to count all the records in a given table. But you might want to count records across multiple subset of data.

For example, as a sales manager, you might want to sum all of the sales of each paper type for each account.

GROUP BY allows you create segments that will be aggregated independently of one another. In other words, **GROUP BY** allows you take the sum of data limited to each account rather than across the entire dataset.

Let's work our way toward this answer by starting with the sums of the quantities of each type of paper.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```
1 SELECT SUM(standard_qty) AS standard_sum,
       SUM(gloss_qty) AS gloss_sum, SUM(poster_qty) AS
       poster_sum
2 FROM orders
```
- Status:** Success!
- Output:** 1 results
- Results:**

standard_sum	gloss_sum	poster_sum
1938346	1013773	723646

Now, we want to create a separate set of sums of each account_id. So, let's add the account_id column into our SELECT statement.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```
1 SELECT account_id, SUM(standard_qty) AS standard_sum,
       SUM(gloss_qty) AS gloss_sum, SUM(poster_qty) AS
       poster_sum
2 FROM orders
```
- Error Message:** column "orders.account_id" must appear in the GROUP BY clause or be used in an aggregate function
- Output:** No results due to error

As you can see, this returns an error. We're effectively collapsing the number of rows returned.

In our previous query, we collapsed all the way down to a single row. The reason we're getting an error here, is the we've included the account_id column but this column isn't being collapsed like the columns that are being aggregated. The query isn't sure whether to sum the account_id as well or whether to make it into a grouping. We have to be explicit about this.

We want to tell the query to aggregate into segments where each segments is one of the values in the account id column. We'll do this using the **GROUP BY** clause.

The screenshot shows a SQL query editor interface. At the top, there are buttons for 'Run' (highlighted in green), 'Format SQL', and 'View History...'. Below the buttons is the SQL code:

```
1 SELECT account_id,
2     SUM(standard_qty) AS standard_sum,
3     SUM(gloss_qty) AS gloss_sum,
4     SUM(poster_qty) AS poster_sum
5 FROM demo.orders
6 GROUP BY account_id
```

On the right side of the editor, a message box displays a green checkmark and the text 'Succeeded in 1s'. Below the editor, there are buttons for 'Export', 'Copy', 'Chart', and 'Pivot'. To the right, it says '350 rows returned' with a small table icon.

	account_id	standard_sum	gloss_sum	poster_sum
1	3601	1818	2332	1436
2	1051	3608	356	156
3	1261	9009	12583	7509
4	2291	4683	5036	2577

You can see that this isn't ordered intuitively, so let's fix that. The **GROUP BY** clause allows goes between the **WHERE** clause if there is one and the **ORDER BY** clause.

The screenshot shows a database query editor interface. On the left, there's a sidebar titled "Input" with a "SCHEMA" section containing tables: accounts, orders, region, sales_reps, and web_events. To the right of the schema is a code editor window displaying the following SQL query:

```

1  SELECT account_id, SUM(standard_qty) AS standard_sum,
2      SUM(gloss_qty) AS gloss_sum, SUM(poster_qty) AS
3          poster_sum
4  FROM orders
5  GROUP BY account_id
6  ORDER BY account_id

```

The status bar at the bottom of the code editor says "Success!" and has a blue "EVALUATE" button. Below the code editor is a table titled "Output" showing 350 results with three columns: account_id, standard_sum, and gloss_sum. The data is as follows:

account_id	standard_sum	gloss_sum	poster_sum
1001	7896	7831	3197
1011	527	14	0
1021	3152	483	175

Whenever there's a field in the **SELECT** statement that's not being aggregated, the query expects it to be in the **GROUP BY** clause. A column that's not aggregated and not in the **GROUP BY** will return the error we saw before. This is pretty important and it's the key to correctly using **GROUP BY** statements. Again, you should always see any columns in the **SELECT** statement that are not being aggregated on, in the **GROUP BY** statement.

GROUP BY II

The key takeaways here:

- **GROUP BY** can be used to aggregate data within subsets of the data. For example, grouping for different accounts, different regions, or different sales representatives.
- Any column in the **SELECT** statement that is not within an aggregator must be in the **GROUP BY** clause.
- The **GROUP BY** always goes between **WHERE** and **ORDER BY**.
- **ORDER BY** works like **SORT** in spreadsheet software.

GROUP BY - Expert Tip

Before we dive deeper into aggregations using **GROUP BY** statements, it is worth noting that SQL evaluates the aggregations before the **LIMIT** clause. If you don't group by any columns, you'll get a 1-row result—no problem there. If you group by a column with enough unique values that it exceeds the **LIMIT** number, the aggregates will be calculated, and then some rows will simply be omitted from the results.

This is actually a nice way to do things because you know you're going to get the correct aggregates. If SQL cuts the table down to 100 rows, then performed the aggregations, your results would be substantially different. The above query's results exceed 100 rows, so it's a perfect example. In the next concept, use the SQL environment to try removing the **LIMIT** and running it again to see what changes.

If we want to segment our data into even more granular chunks, we can GROUP BY multiple columns. Imagine yourself as a marketing manager at Parch & Posey, trying to understand how each account interacted with various advertising channels.

Which channels are driving traffic and leading to purchases? Are we investing in channels that aren't worth the cost ? how much traffic are we obtaining from each channel ?

One way we might begin to look into these questions, is to count up all of the events for each channel, for each account id.

The screenshot shows a database interface with the following details:

Input Schema:

- accounts
- orders
- region
- sales_reps
- web_events

Query:

```
1 SELECT account_id, channel, COUNT(id) AS events
2 FROM web_events w
3 GROUP BY 1,2
4 ORDER BY 1,2
```

Status: Success!

EVALUATE button

Output: 1509 results

account_id	channel	events
1001	adwords	5
1001	banner	3
1001	direct	22

It looks like the really useful information here is in the events column. Let's reorder this to highlight the highest volume channels for each account.

The screenshot shows a database interface with the following details:

Input Schema:

- accounts
- orders
- region
- sales_reps
- web_events

Query:

```
1 SELECT account_id, channel, COUNT(id) AS events
2 FROM web_events w
3 GROUP BY 1,2
4 ORDER BY 1,2 desc
```

Status: Success!

EVALUATE button

Output: 1509 results

account_id	channel	events
1001	twitter	1
1001	organic	6
1001	facebook	2

You'll notice that account id comes before events in the order by clause. Here in the results, you can see that these are ordered first by account ID, then by events within the account ID.

Quiz: GROUP BY

GROUP BY Note

Now that you have been introduced to **JOINS**, **GROUP BY**, and aggregate functions, the real power of **SQL** starts to come to life. Try some of the below to put your skills to the test!

Questions: GROUP BY

Use the **SQL** environment below to assist with answering the following questions. Whether you get stuck or you just want to double check your solutions, my answers can be found at the top of the next concept.

One part that can be difficult to recognize is when it might be easiest to use an aggregate or one of the other SQL functionalities. Try some of the below to see if you can differentiate to find the easiest solution.

1. Which **account** (by name) placed the earliest order? Your solution should have the **account name** and the **date** of the order.

The screenshot shows a SQL environment interface. The top bar has tabs for 'Input' (selected), 'HISTORY', and 'MENU'. In the 'Input' tab, there is a schema dropdown showing 'accounts', 'orders', 'region', 'sales_reps', and 'web_events'. Below the dropdown is a code editor containing the following SQL query:

```
1 SELECT a.name, o.occurred_at
2 FROM accounts a
3 JOIN orders o
4 ON a.id = o.account_id
5 ORDER BY occurred_at
6 LIMIT 1;
```

Below the code editor, a message says 'Success!' and there is a blue 'EVALUATE' button. Under the 'Output' section, it says '1 results' and shows a table with two columns: 'name' and 'occurred_at'. The single result is 'DISH Network' in the 'name' column and '2013-12-04T04:22:44.000Z' in the 'occurred_at' column. There is also a link to 'Download CSV'.

2. Find the total sales in **usd** for each account. You should include two columns - the total sales for each company's orders in **usd** and the company **name**.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```

1  SELECT a.name, SUM(total_amt_usd) total_sales
2  FROM orders o
3  JOIN accounts a
4  ON a.id = o.account_id
5  GROUP BY a.name;
    
```

- Status:** Success!
- EVALUATE** button.
- Output:** 350 results
- Download CSV** button.
- Table Data:**

name	total_sales
Boeing	65091.39
Western Digital	269155.34
Sysco	278575.64
Southern	42881.21

3. Via what **channel** did the most recent (latest) **web_event** occur, which **account** was associated with this **web_event**? Your query should return only three values - the **date**, **channel**, and **account name**.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```

1  SELECT w.occurred_at, w.channel, a.name
2  FROM web_events w
3  JOIN accounts a
4  ON w.account_id = a.id
5  ORDER BY w.occurred_at DESC
6  LIMIT 1;
    
```

- Status:** Success!
- EVALUATE** button.
- Output:** 1 results
- Download CSV** button.
- Table Data:**

occurred_at	channel	name
2017-01-01T23:51:09.000Z	organic	Molina Healthcare

4. Find the total number of times each type of **channel** from the **web_events** was used. Your final table should have two columns - the **channel** and the number of times the channel was used.

Input

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT w.channel, COUNT(*)
2 FROM web_events w
3 GROUP BY w.channel
```

Success!

EVALUATE

Output 6 results

Download CSV

channel	count
adwords	906
direct	5298
banner	476

5. Who was the **primary contact** associated with the earliest **web_event**?

Input

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT a.primary_poc
2 FROM web_events w
3 JOIN accounts a
4 ON a.id = w.account_id
5 ORDER BY w.occurred_at
6 LIMIT 1;
```

Success!

EVALUATE

Output 1 results

Download CSV

primary_poc
Leana Hawker

6. What was the smallest order placed by each **account** in terms of **total usd**. Provide only two columns - the account **name** and the **total usd**. Order from smallest dollar amounts to largest.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```

1  SELECT a.name, MIN(total_amt_usd) smallest_order
2  FROM accounts a
3  JOIN orders o
4  ON a.id = o.account_id
5  GROUP BY a.name
6  ORDER BY smallest_order;
    
```

- Status:** Success!
- Output:** 350 results
- Download CSV** button
- Table Data:**

name	smallest_order
Gilead Sciences	0.00
FedEx	0.00
Navistar International	0.00
General Mills	0.00

7. Find the number of **sales reps** in each region. Your final table should have two columns - the **region** and the number of **sales_reps**. Order from fewest reps to most reps.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```

1  SELECT r.name, COUNT(*) num_reps
2  FROM region r
3  JOIN sales_reps s
4  ON r.id = s.region_id
5  GROUP BY r.name
6  ORDER BY num_reps
    
```

- Status:** Success!
- Output:** 4 results
- Download CSV** button
- Table Data:**

name	num_reps
Midwest	9
Southeast	10
West	10
Northeast	21

Quiz: GROUP BY Part II

Questions: GROUP BY Part II

Use the **SQL** environment below to assist with answering the following questions. Whether you get stuck or you just want to double check your solutions, my answers can be found at the top of the next concept.

1. For each account, determine the average amount of each type of paper they purchased across their orders. Your result should have four columns - one for the account **name** and one for the average quantity purchased for each of the paper types for each account.

The screenshot shows a SQL environment interface with the following components:

- Input:** A sidebar on the left listing available schemas: accounts, orders, region, sales_reps, and web_events. The 'accounts' schema is currently selected.
- Query Editor:** The main area contains a SQL query:

```
1 SELECT a.name, AVG(o.standard_qty) avg_stand,
2     AVG(o.gloss_qty) avg_gloss, AVG(o.poster_qty) avg_post
3 FROM accounts a
4 JOIN orders o
5 ON a.id = o.account_id
6 GROUP BY a.name;
```
- Status:** Below the query editor, a green message says "Success!"
- Evaluate:** A blue button labeled "EVALUATE" is located to the right of the status message.
- Output:** Below the input section, it says "Output 350 results".
- Download CSV:** A link to download the results as a CSV file.
- Results Table:** A table showing the output of the query. The columns are "name", "avg_stand", "avg_gloss", and "avg_post". The rows are:

name	avg_stand	avg_gloss	avg_post
Boeing	360.266666666666666667	222.3333333333333333	1
Western Digital	192.7846153846153846	239.8307692307692308	1
Sysco	191.0588235294117647	231.1764705882352941	1
Southern	347.500000000000000000	27.750000000000000000	2

2. For each account, determine the average amount spent per order on each paper type. Your result should have four columns - one for the account **name** and one for the average amount spent on each paper type.

Input

```

1  SELECT a.name, AVG(o.standard_amt_usd) avg_stand,
2      AVG(o.gloss_amt_usd) avg_gloss, AVG(o.poster_amt_usd)
3      avg_post
4  FROM accounts a
5  JOIN orders o
6  ON a.id = o.account_id
7  GROUP BY a.name;
  
```

Output 350 results

name	avg_stand	avg_gloss	avg_post
Boeing	1797.7306666666666667	1665.2766666666666667	8
Western Digital	961.9952307692307692	1796.3324615384615385	1
Sysco	953.3835294117647059	1731.5117647058823529	1
Southern	1734.0250000000000000	207.8475000000000000	2

3. Determine the number of times a particular **channel** was used in the **web_events** table for each **sales rep**. Your final table should have three columns - the **name of the sales rep**, the **channel**, and the number of occurrences. Order your table with the highest number of occurrences first.

Input

```

1  SELECT s.name, w.channel, COUNT(*) num_events
2  FROM accounts a
3  JOIN web_events w
4  ON a.id = w.account_id
5  JOIN sales_reps s
6  ON s.id = a.sales_rep_id
7  GROUP BY s.name, w.channel
8  ORDER BY num_events DESC;
  
```

Output 295 results

name	channel	num_events
Earlie Schleusner	direct	234
Vernita Plump	direct	232
Moon Torian	direct	194
Georgianna Chisholm	direct	188

4. Determine the number of times a particular **channel** was used in the **web_events** table for each **region**. Your final table should have three columns - the **region name**, the **channel**, and the number of occurrences. Order your table with the highest number of occurrences first.

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1 SELECT r.name, w.channel, COUNT(*) num_events
2 FROM accounts a
3 JOIN web_events w
4 ON a.id = w.account_id
5 JOIN sales_reps s
6 ON s.id = a.sales_rep_id
7 JOIN region r
8 ON r.id = s.region_id
9 GROUP BY r.name, w.channel
10 ORDER BY num_events DESC;
```

Success! EVALUATE

Output 24 results [Download CSV](#)

name	channel	num_events
Northeast	direct	1800
Southeast	direct	1548
West	direct	1254
Midwest	direct	696

DISTINCT

DISTINCT is always used in **SELECT** statements, and it provides the unique rows for all columns written in the **SELECT** statement. Therefore, you only use **DISTINCT** once in any particular **SELECT** statement.

You could write:

```
SELECT DISTINCT column1, column2, column3  
FROM table1;
```

which would return the unique (or **DISTINCT**) rows across all three columns.

You would **not** write:

```
SELECT DISTINCT column1, DISTINCT column2, DISTINCT column3  
FROM table1;
```

You can think of **DISTINCT** the same way you might think of the statement "unique".

DISTINCT - Expert Tip

It's worth noting that using **DISTINCT**, particularly in aggregations, can slow your queries down quite a bit.

You can think of **DISTINCT** this way, if you want to GROUP BY some columns but you don't necessarily want to include any aggregations, you can use **DISTINCT** instead.

As an example, let's revisit the count of events by channel by account we looked at in the GROUP BY segment.

Input

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT account_id, channel, COUNT(id) AS events
2 FROM web_events w
3 GROUP BY 1,2
4 ORDER BY 1,2 desc
```

Success!

EVALUATE

Output 1509 results

account_id	channel	events
1001	twitter	1
1001	organic	6
1001	facebook	2

Notice the row count, 1509. If we get rid of the events column, you can see that the resulting query returns basically the same results with the same row count.

Input

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT account_id, channel
2 FROM web_events w
3 GROUP BY 1,2
4 ORDER BY 1
```

Success!

EVALUATE

Output 1509 results

account_id	channel
1001	facebook
1001	direct
1001	twitter

And if we run that query again with **DISTINCT**, you can see that the results are the same.

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1 SELECT DISTINCT account_id, channel
2 FROM web_events w
3 GROUP BY 1,2
4 ORDER BY 1
```

Success! EVALUATE

Output 1509 results

account_id	channel
1001	adwords
1001	banner
1001	direct

Of course, they don't include the COUNT anymore because we've removed it from our select statement. Since it's no longer in our select statement, and since we're no longer doing an aggregation, we don't need to GROUP BY clause either.

Quiz: DISTINCT

Questions: DISTINCT

Use the **SQL** environment below to assist with answering the following questions. Whether you get stuck or you just want to double check your solutions, my answers can be found at the top of the next concept.

1. Use **DISTINCT** to test if there are any accounts associated with more than one region.

The screenshot shows a SQL environment interface. On the left, there's a schema browser with tables: accounts, orders, region, sales_reps, and web_events. The 'accounts' table is currently selected. On the right, the code editor contains the following SQL query:

```
1 SELECT a.id as "account id", r.id as "region id",
2 a.name as "account name", r.name as "region name"
3 FROM accounts a
4 JOIN sales_reps s
5 ON s.id = a.sales_rep_id
6 JOIN region r
7 ON r.id = s.region_id;
```

The status bar at the bottom of the code editor says "Success!" and has a blue "EVALUATE" button. Below the code editor is a table titled "Output" showing 351 results. The table has four columns: account id, region id, account name, and region name. The data is as follows:

account id	region id	account name	region name
1691	1	Johnson Controls	Northeast
1661	1	American Airlines Group	Northeast
1631	1	Ingram Micro	Northeast

In another way

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1   SELECT DISTINCT id, name
2   FROM accounts;

```

Success!

EVALUATE

Output 351 results

Download CSV

id	name
1621	HCA Holdings
2201	Plains GP Holdings
1171	Amazon.com

2. Have any **sales reps** worked on more than one account?

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1   SELECT s.id, s.name, COUNT(*) num_accounts
2   FROM accounts a
3   JOIN sales_reps s
4   ON s.id = a.sales_rep_id
5   GROUP BY s.id, s.name
6   ORDER BY num_accounts;

```

Success!

EVALUATE

Output 50 results

Download CSV

id	name	num_accounts
321660	Silvana Virden	3
321570	Shawanda Selke	3
321790	Cordell Rieder	3
321650	Akilah Drinkard	3

In another way

The screenshot shows a database interface with the following details:

Input:

- SCHEMA:** A dropdown menu showing options: accounts, orders, region, sales_reps, web_events.
- Query:** A code editor containing the following SQL query:

```
1  SELECT DISTINCT id, name
2  FROM sales_reps;
```
- Status:** A message "Success!" is displayed below the query.
- EVALUATE:** A blue button to execute the query.

Output: 50 results

id	name
321770	Kathleen Lalonde
321620	Retha Sears
321960	Maryanna Fiorentino
321690	Gianna Dossey

Download CSV: A link to download the results as a CSV file.

HAVING

HAVING - Expert Tip

HAVING is the “clean” way to filter a query that has been aggregated, but this is also commonly done using a [subquery](#). Essentially, any time you want to perform a **WHERE** on an element of your query that was created by an aggregate, you need to use **HAVING** instead.

Example :

imagine yourself as an account manager at Parch & Posey, working with the company's largest accounts. you might want to identify the total sales in US dollars for account with over \$250,000 in sales, to better understand the profession of revenue that comes from these large accounts. To get this list, let's first get the sum sales for each account. We'll order them in descending order so that we can see all the highest value accounts at the top of the result set.

The screenshot shows a SQL query editor with the following details:

- Toolbar: Run, Limit 100, Format SQL, View History...
- Query Text:

```
1 SELECT account_id,
2        SUM(total_amt_usd) AS sum_total_amt_usd
3     FROM demo.orders
4    GROUP BY 1
5 ORDER BY 2 DESC
```
- Status Bar: ✓ Succeeded in 948ms
- Result Table:

account_id	sum_total_amt_usd
4211	382873.3
4151	345618.5900000001
1301	326819.4800000004
1871	300694.7899999999
- Bottom Navigation: Export, Copy, Chart, Pivot, 100 rows returned

Now let's filter down to just the accounts with more than \$250,000 in sales.

The screenshot shows a SQL editor interface with a dark theme. At the top, there are buttons for 'Run', 'Limit 100', 'Format SQL', and 'View History...'. The SQL query is:

```
1 SELECT account_id,
2     SUM(total_amt_usd) AS sum_total_amt_usd
3 FROM demo.orders
4 WHERE SUM(total_amt_usd) >= 250000
5 GROUP BY 1
6 ORDER BY 2 DESC
```

A yellow warning box at the bottom right says 'There was a problem with your query'. Below the editor, a message says 'Looks like something went wrong with your query. [Get help here.](#)' followed by the error details:

aggregate functions are not allowed in WHERE
Position: 94

FROM demo.orders
WHERE SUM(total_amt_usd) >= 250000
^
GROUP BY 1

You'll notice that the WHERE clause isn't exactly built for this. The WHERE clause won't work for this because it doesn't allow you to filter on aggregate columns. That's where the HAVING clause comes in.

This filter query down to just the account IDs with more than \$250,000 in total sales.

The screenshot shows a SQL editor interface with a dark theme. At the top, there are buttons for 'Run', 'Limit 100', 'Format SQL', and 'View History...'. The SQL query is:

```
1 SELECT account_id,
2     SUM(total_amt_usd) AS sum_total_amt_usd
3 FROM demo.orders
4 GROUP BY 1
5 HAVING SUM(total_amt_usd) >= 250000
```

A green success message at the bottom right says 'Succeeded in 1s'. Below the editor, there is a table with the results:

	account_id	sum_total_amt_usd
1	4211	382873.3
2	3411	291047.2500000006
3	4251	255319.1800000002
4	2931	269155.3399999997
5	1871	300694.7899999999

One thing to keep in mind is that this is really only useful when grouping by one or more columns. if you're aggregating across the entire dataset, the output is only one line anyway, so there's no need to filter beyond that.

Quiz : HAVING

Often there is confusion about the difference between **WHERE** and **HAVING**. Select all the statements that are true regarding **HAVING** and **WHERE** statements.

- WHERE** subsets the returned data based on a logical condition.
- WHERE** appears after the **FROM**, **JOIN**, and **ON** clauses, but before **GROUP BY**.
- HAVING** appears after the **GROUP BY** clause, but before the **ORDER BY** clause.
- HAVING** is like **WHERE**, but it works on logical statements involving aggregations.

Questions: HAVING

Use the **SQL** environment below to assist with answering the following questions. Whether you get stuck or you just want to double check your solutions, my answers can be found at the top of the next concept.

1. How many of the **sales reps** have more than 5 accounts that they manage?

Input

```
1  SELECT s.id, s.name, COUNT(*) num_accounts
2  FROM accounts a
3  JOIN sales_reps s
4  ON s.id = a.sales_rep_id
5  GROUP BY s.id, s.name
6  HAVING COUNT(*) > 5
7  ORDER BY num_accounts;
```

Success!

EVALUATE

Output 34 results

Download CSV

id	name	num_accounts
321500	Samuel Racine	6
321510	Eugena Esser	6
321560	Elba Felder	6
321590	Necole Victory	6

and technically, we can get this using a **SUBQUERY** as shown below. This same logic can be used for the other queries, but this will not be shown.

Input

```
1  SELECT COUNT(*) num_reps_above5
2  FROM(SELECT s.id, s.name, COUNT(*) num_accounts
3  FROM accounts a
4  JOIN sales_reps s
5  ON s.id = a.sales_rep_id
6  GROUP BY s.id, s.name
7  HAVING COUNT(*) > 5
8  ORDER BY num_accounts) AS Table1;
```

Success!

EVALUATE

Output 1 results

Download CSV

num_reps_above5
34

2. How many **accounts** have more than 20 orders?

Input

HISTORY ▾ MENU ▾

SCHEMA	▼
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1 SELECT a.id, a.name, COUNT(*) num_orders
2 FROM accounts a
3 JOIN orders o
4 ON a.id = o.account_id
5 GROUP BY a.id, a.name
6 HAVING COUNT(*) > 20
7 ORDER BY num_orders;
```

Success!

EVALUATE

Output 120 results

Download CSV

id	name	num_orders
2841	Performance Food Group	21
4171	Thrivent Financial for Lutherans	21
1321	Anthem	21
2571	Jabil Circuit	22

3. Which account has the most orders?

Input

HISTORY ▾ MENU ▾

SCHEMA	▼
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1 SELECT a.id, a.name, COUNT(*) num_orders
2 FROM accounts a
3 JOIN orders o
4 ON a.id = o.account_id
5 GROUP BY a.id, a.name
6 ORDER BY num_orders DESC
7 LIMIT 1;
```

Success!

EVALUATE

Output 1 results

Download CSV

id	name	num_orders
3411	Leucadia National	71

4. How many accounts spent more than 30,000 usd total across all orders?

Input

HISTORY ▾ MENU ▾

SCHEMA

accounts

orders

region

sales_reps

web_events

```
1 SELECT a.id, a.name, SUM(o.total_amt_usd) total_spent
2 FROM accounts a
3 JOIN orders o
4 ON a.id = o.account_id
5 GROUP BY a.id, a.name
6 HAVING SUM(o.total_amt_usd) > 30000
7 ORDER BY total_spent;
```

Success!

EVALUATE

Output 204 results

Download CSV

id	name	total_spent
1661	American Airlines Group	30083.18
1431	PepsiCo	30095.72
3661	Group 1 Automotive	30708.92
1141	Costco	30741.01

5. How many accounts spent less than 1,000 usd total across all orders?

Input

HISTORY ▾ MENU ▾

SCHEMA

accounts

orders

region

sales_reps

web_events

```
1 SELECT a.id, a.name, SUM(o.total_amt_usd) total_spent
2 FROM accounts a
3 JOIN orders o
4 ON a.id = o.account_id
5 GROUP BY a.id, a.name
6 HAVING SUM(o.total_amt_usd) < 1000
7 ORDER BY total_spent;
```

Success!

EVALUATE

Output 3 results

Download CSV

id	name	total_spent
1901	Nike	390.25
1671	Delta Air Lines	859.64
4321	Level 3 Communications	881.73

6. Which account has spent the most with us?

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1 SELECT a.id, a.name, SUM(o.total_amt_usd) total_spent
2 FROM accounts a
3 JOIN orders o
4 ON a.id = o.account_id
5 GROUP BY a.id, a.name
6 ORDER BY total_spent DESC
7 LIMIT 1;
```

Success! EVALUATE

Output 1 results [Download CSV](#)

id	name	total_spent
4211	EOG Resources	382873.30

7. Which account has spent the least with us?

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1 SELECT a.id, a.name, SUM(o.total_amt_usd) total_spent
2 FROM accounts a
3 JOIN orders o
4 ON a.id = o.account_id
5 GROUP BY a.id, a.name
6 ORDER BY total_spent
7 LIMIT 1
```

Success! EVALUATE

Output 1 results [Download CSV](#)

id	name	total_spent
1901	Nike	390.25

8. Which accounts used Facebook as a **channel** to contact customers more than 6 times?

Input

SCHEMA	accounts	orders	region	sales_reps	web_events

```

1  SELECT a.id, a.name, w.channel, COUNT(*) use_of_channel
2  FROM accounts a
3  JOIN web_events w
4  ON a.id = w.account_id
5  GROUP BY a.id, a.name, w.channel
6  HAVING COUNT(*) > 6 AND w.channel = 'facebook'
7  ORDER BY use_of_channel;

```

Success! **EVALUATE**

Output 46 results [Download CSV](#)

id	name	channel	use_of_channel
3261	Farmers Insurance Exchange	facebook	7
3231	Parker-Hannifin	facebook	7
4241	Laboratory Corp. of America	facebook	7
1741	Honeywell International	facebook	7

9. Which account used facebook most as a **channel**?

Input

SCHEMA	accounts	orders	region	sales_reps	web_events

```

1  SELECT a.id, a.name, w.channel, COUNT(*) use_of_channel
2  FROM accounts a
3  JOIN web_events w
4  ON a.id = w.account_id
5  WHERE w.channel = 'facebook'
6  GROUP BY a.id, a.name, w.channel
7  ORDER BY use_of_channel DESC
8  LIMIT 1;

```

Success! **EVALUATE**

Output 1 results [Download CSV](#)

id	name	channel	use_of_channel
1851	Gilead Sciences	facebook	16

10. Which channel was most frequently used by most accounts?

Input

HISTORY ▾ MENU ▾

SCHEMA	▼
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1  SELECT a.id, a.name, w.channel, COUNT(*) use_of_channel
2  FROM accounts a
3  JOIN web_events w
4  ON a.id = w.account_id
5  GROUP BY a.id, a.name, w.channel
6  ORDER BY use_of_channel DESC
7  LIMIT 10;
```

Success!

EVALUATE

Output 10 results

Download CSV

id	name	channel	use_of_channel
3411	Leucadia National	direct	52
1601	New York Life Insurance	direct	51
2731	Colgate-Palmolive	direct	51
2051	Philip Morris International	direct	49

DATE Functions

GROUPing BY a date column is not usually very useful in SQL, as these columns tend to have transaction data down to a second. Keeping date information at such a granular data is both a blessing and a curse, as it gives really precise information (a blessing), but it makes grouping information together directly difficult (a curse).

Lucky for us, there are a number of built in SQL functions that are aimed at helping us improve our experience in working with dates.

Here we saw that dates are stored in year, month, day, hour, minute, second, which helps us in truncating. In the next concept, you will see a number of functions we can use in SQL to take advantage of this functionality.

By now, you might have noticed that dates are a bit hard to work with. Aggregating by data fields, in particular, doesn't work in a practical way. It treats each time stamp is unique. When it would be more practical to round to the nearest day, week or month and aggregate across that period. Take for example, this sum of standard paper quantities by time period.

The screenshot shows a SQL query editor interface. On the left, there's a schema browser with tables: accounts, orders, region, sales_reps, and web_events. The 'orders' table is currently selected. On the right, the query pane contains the following SQL code:

```
1  SELECT occurred_at, SUM(standard_qty) AS
2    standard_qty_sum
3  FROM orders o
4  GROUP BY occurred_at
5  ORDER BY occurred_at
```

The status bar at the bottom of the query pane says "Success!". Below the query pane is the output pane, which displays the results of the query:

occurred_at	standard_qty_sum
2013-12-04T04:22:44.000Z	0
2013-12-04T04:45:54.000Z	490
2013-12-04T04:53:25.000Z	528
2013-12-05T20:29:16.000Z	0

you can see in the results here that this really isn't any more useful than booking at the raw data. This aggregates from six 6,912 rows in the raw data down to 6,908. Almost all of the date in this table are unique. good news though, there are plenty of special functions to make dates easier to work with. But before we dig into date and time functions, let's take a look at how data stored.

If you live in the United States, you're probably used to seeing dates formatted as month-month, day-day, year-year-year-year, or similar month first format. This isn't necessarily better or rows, it's just different. Databases do it yet another way. Ordering from least to most granular part of the data, year- year- year, month-month, day-day. This is a very specific utility, and bases of that date stored alphabetically are also in chronological order. In other words, date ordering is the same whether you think of them as dates or as bits of text.

Databases	In the U.S.	In the rest of the world
YYYY MM DD	MM DD YY	DD MM YY
2015-09-21	03-19-2016	08-12-2016
2016-03-19	09-21-2015	10-10-2017
2016-12-08	10-10-2017	19-03-2016
2017-10-10	10-10-2017	21-09-2015

Sorted **oldest** to **newest** based on alphabetical sorting

With this example here, we can see that the year first way the databases store dates, is ideal for sorting the way we'll want to retrieve this information in the future. Whether we want the most recent or oldest information, day first and month first date formats sort of funny ways that don't make a ton of sense. Another benefit is the date can easily be truncated in order to group them for analyses.

Take the date, **2017 - 04 - 01 12:15:01**. If we want to group that with other events that occurred on the same day, we can't do that with the date in its current format. Grouping now or group by every event that occurred in April 1st, at 12:15 and 1 second. That won't help us very much.

DATE Functions II

The first function you are introduced to in working with dates is **DATE_TRUNC**.

DATE_TRUNC allows you to truncate your date to a particular part of your date-time column. Common truncations are **day**, **month**, and **year**.

DATE_PART can be useful for pulling a specific portion of a date, but notice pulling **month** or day of the week (**dow**) means that you are no longer keeping the years in order. Rather you are grouping for certain components regardless of which year they belonged in.

For additional functions you can use with dates, but the **DATE_TRUNC** and **DATE_PART** functions definitely give you a great start!

You can reference the columns in your select statement in **GROUP BY** and **ORDER BY** clauses with numbers that follow the order they appear in the select statement. For example

```
SELECT standard_qty, COUNT(*)
```

```
FROM orders
```

GROUP BY 1 (this 1 refers to standard_qty since it is the first of the columns included in the select statement)

ORDER BY 1 (this 1 refers to standard_qty since it is the first of the columns included in the select statement)

For example, in order to group by day, we'll need to adjust all the times on April 1st 2017 to read:

2017-04-01 00:00:00. That way when we group by date, we get every event that occurred for all hours, minutes and seconds of April 1st. They'll all be grouped together into the same grouping.

We can do this using the date trunc function.

Let's start by grouping by occurred at without any truncation.

The screenshot shows a data visualization interface with the following components:

- Input:** A sidebar listing databases and tables: accounts, orders, region, sales_reps, and web_events.
- Query Editor:** A code editor containing the following SQL query:


```

1 SELECT occurred_at, SUM(standard_qty) AS
2   standard_qty_sum
3   FROM orders o
4   GROUP BY occurred_at
5   ORDER BY occurred_at
      
```
- Status:** A message "Success!" is displayed below the query.
- Evaluate Button:** A blue button labeled "EVALUATE" is located at the bottom right of the code editor.
- Output:** A table showing the results of the query:

occurred_at	standard_qty_sum
2013-12-04T04:22:44.000Z	0
2013-12-04T04:45:54.000Z	490
2013-12-04T04:53:25.000Z	528
2013-12-05T20:29:16.000Z	0

As you can see, this doesn't help us much at all. If we replace each instance of occurred at with a truncated version, we'll get result set that sums the quantities of standard paper by day. It's important to group by the same metric that's included in the select statement to assure your results are consisting. In some databases, it's possible to group by a year but truncate on day in the select statement.

The screenshot shows a database query tool with the following interface elements:

- Toolbar:** Buttons for Run, Limit 100, Format SQL, and View History...
- Query Editor:** A code editor containing the following SQL query:


```

1 SELECT DATE_TRUNC('day',occurred_at) AS day,
2   SUM(standard_qty) AS standard_qty_sum
3   FROM demo.orders
4   GROUP BY DATE_TRUNC('day',occurred_at)
5   ORDER BY DATE_TRUNC('day',occurred_at)
      
```
- Status:** A message "✓ Succeeded in 1s" is displayed in the bottom right corner.
- Results:** A table showing the results of the query:

day	standard_qty_sum
2013-12-04 00:00:00	1018
2013-12-05 00:00:00	492
2013-12-06 00:00:00	1692
2013-12-08 00:00:00	3877

This is something most people only do by accident as a provider results that are confusing and in many cases wrong for the type of question you're trying to answer. the easiest way to make sure you correctly is to use column numbers instead of retyping the exact functions. Date trunc can be used to aggregate at very granular levels like second. That might be useful if you're working with server logs or many events happen in give second.

RESULT	INPUT
2017-04-01 12:15:01	DATE_TRUNC ('second', 2017-04-01 12:15:01)
2017-04-01 00:00:00	DATE_TRUNC ('day', 2017-04-01 12:15:01)
2017-04-01 00:00:00	DATE_TRUNC ('month', 2017-04-01 12:15:01)
2017-01-01 00:00:00	DATE_TRUNC ('year', 2017-04-01 12:15:01)

most of the time though, you'll use this to aggregate at intervals that make sense from a business perspective: day, week, month, quarter and year. Here, we can see different truncations. Notice that if our month or day are 01, There's no change made to these values, as you can see.

there are some cases where you might want to just pull out again part of the day. for example if you want to know what day of the week Parch & Posey's website sees the most traffic, you wouldn't want to use date trunk. To get the way of week, you'd have to use date part. Date part allows you to pull the part of the date that you're interested in. but notice that regardless of year, a date part would provide the same month for an event that happens in April 2016 and April 2017 where a date trunk would differentiate these events.

RESULT	INPUT
1	DATE_PART ('second', 2017-04-01 12:15:01)
1	DATE_PART ('day', 2017-04-01 12:15:01)
4	DATE_PART ('month', 2017-04-01 12:15:01)
2017	DATE_PART ('year', 2017-04-01 12:15:01)

Let's explore this example using Parch & Posey's date. On what day of the week are the most sales made?

Let's start by figuring out the day of the week for each one. DOW stands for day of week and returns a value from zero to six, where zero is Sunday and six is Saturday.

Now that we have this column, we can aggregate to figure out the day with the most reams of paper sold. We'll order it by the sum descending order so that the day with the most sales will be at the top of the result set.

The screenshot shows a SQL query being run in a database interface. The query is:

```
1 SELECT DATE_PART('dow', occurred_at) AS day_of_week,
2        SUM(total) AS total_qty
3    FROM demo.orders
4   GROUP BY 1
5 ORDER BY 2 DESC;
```

The results table has two columns: `day_of_week` and `total_qty`. The data is:

day_of_week	total_qty
0	559873
5	536776
6	521813
1	518016

At the bottom, it says "7 rows returned".

Look like the most paper was ordered on Sunday and the last paper was ordered on Thursday.

Quiz: DATE Functions

Questions: Working With DATEs

Use the **SQL** environment below to assist with answering the following questions. Whether you get stuck or you just want to double check your solutions, my answers can be found at the top of the next concept.

1. Find the sales in terms of total dollars for all orders in each **year**, ordered from greatest to least. Do you notice any trends in the yearly sales totals?

The screenshot shows a SQL environment interface. On the left, there's a sidebar titled "Input" with a dropdown menu for "SCHEMA" containing "accounts", "orders", "region", "sales_reps", and "web_events". The main area contains a code editor with the following SQL query:

```
1  SELECT DATE_PART('year', occurred_at) ord_year,
2      SUM(total_amt_usd) total_spent
3  FROM orders
4  GROUP BY 1
5  ORDER BY 2 DESC;
```

Below the code editor, a message says "Success!" and there's a blue "EVALUATE" button. Underneath, the "Output" section shows a table with 5 results:

ord_year	total_spent
2016	12864917.92
2015	5752004.94
2014	4069106.54

At the bottom right of the output table, there's a link "Download CSV".

When we look at the yearly totals, you might notice that 2013 and 2017 have much smaller totals than all other years. If we look further at the monthly data, we see that for **2013** and **2017** there is only one month of sales for each of these years (12 for 2013 and 1 for 2017). Therefore, neither of these are evenly represented. Sales have been increasing year over year, with 2016 being the largest sales to date. At this rate, we might expect 2017 to have the largest sales.

2. Which **month** did Parch & Posey have the greatest sales in terms of total dollars?

Are all months evenly represented by the dataset?

In order for this to be 'fair', we should remove the sales from 2013 and 2017. For the same reasons as discussed above.

Input

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT DATE_PART('month', occurred_at) ord_month,  
2     SUM(total_amt_usd) total_spent  
3 FROM orders  
4 WHERE occurred_at BETWEEN '2014-01-01' AND '2017-01-01'  
5 GROUP BY 1  
6 ORDER BY 2 DESC;
```

Success!

EVALUATE

Output 12 results

Download CSV

ord_month	total_spent
12	2752080.98
10	2427505.97
11	2390033.75
9	2017216.88

The greatest sales amounts occur in December (12).

3. Which **year** did Parch & Posey have the greatest sales in terms of total number of orders? Are all years evenly represented by the dataset?

Input

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT DATE_PART('year', occurred_at) ord_year,  
2     COUNT(*) total_sales  
3 FROM orders  
4 GROUP BY 1  
5 ORDER BY 2 DESC;
```

Success!

EVALUATE

Output 5 results

Download CSV

ord_year	total_sales
2016	3757
2015	1725
2014	1306

Again, 2016 by far has the most amount of orders, but again 2013 and 2017 are not evenly represented to the other years in the dataset.

4. Which **month** did Parch & Posey have the greatest sales in terms of total number of orders? Are all months evenly represented by the dataset?

Input

SCHEMA

accounts

orders

region

sales_reps

web_events

1 SELECT DATE_PART('month', occurred_at) ord_month,
2 COUNT(*) total_sales
3 FROM orders
4 WHERE occurred_at BETWEEN '2014-01-01' AND '2017-01-01'
5 GROUP BY 1
6 ORDER BY 2 DESC;

Success!

EVALUATE

Output 12 results

Download CSV

ord_month	total_sales
12	783
11	713
10	675
8	603

December still has the most sales, but interestingly, November has the second most sales (but not the most dollar sales. To make a fair comparison from one month to another 2017 and 2013 data were removed.

5. In which **month** of which **year** did **Walmart** spend the most on gloss paper in terms of dollars?

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1 SELECT DATE_TRUNC('month', o.occurred_at) ord_date,  
2     SUM(o.gloss_amt_usd) tot_spent  
3 FROM orders o  
4 JOIN accounts a  
5 ON a.id = o.account_id  
6 WHERE a.name = 'Walmart'  
7 GROUP BY 1  
8 ORDER BY 2 DESC  
9 LIMIT 1;
```

Success! EVALUATE

Output 1 results [Download CSV](#)

ord_date	tot_spent
2016-05-01T00:00:00.000Z	9257.64

May 2016 was when Walmart spent the most on gloss paper.

CASE Statements

CASE - Expert Tip

- The CASE statement always goes in the SELECT clause.
- CASE must include the following components: WHEN, THEN, and END. ELSE is an optional component to catch cases that didn't meet any of the other previous CASE conditions.
- You can make any conditional statement using any conditional operator (like **WHERE**) between WHEN and THEN. This includes stringing together multiple conditional statements using AND and OR.
- You can include multiple WHEN statements, as well as an ELSE statement again, to deal with any unaddressed conditions.

Example

In a quiz question in the previous Basic SQL lesson, you saw this question:

1. Create a column that divides the `standard_amt_usd` by the `standard_qty` to find the unit price for standard paper for each order. Limit the results to the first 10 orders, and include the `id` and `account_id` fields. **NOTE - you will be thrown an error with the correct solution to this question. This is for a division by zero. You will learn how to get a solution without an error to this query when you learn about CASE statements in a later section.**

Let's see how we can use the **CASE** statement to get around this error.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```

1  SELECT id, account_id, standard_amt_usd/standard_qty AS
2    unit_price
3   FROM orders
4   LIMIT 10;

```

- Status:** Success!
- Output:** 10 results

	id	account_id	unit_price
1		1001	4.990000000000000
2		1001	4.990000000000000
3		1001	4.990000000000000
4		1001	4.990000000000000

Now, let's use a **CASE** statement. This way any time the **standard_qty** is zero, we will return 0, and otherwise we will return the **unit_price**.

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```

1  SELECT account_id, CASE WHEN standard_qty = 0 OR
2    standard_qty IS NULL THEN 0
3    ELSE
4      standard_amt_usd/standard_qty END AS unit_price
5   FROM orders
6   LIMIT 10;

```

- Status:** Success!
- Output:** 10 results

	account_id	unit_price
1	1001	4.990000000000000
2	1001	4.990000000000000
3	1001	4.990000000000000
4	1001	4.990000000000000

Let's see how we can use the **CASE** statement to get around this error.

Now the first part of the statement will catch any of those division by zero values that were causing the error, and the other components will compute the division as necessary. You will notice, we essentially charge all of our accounts 4.99 for standard paper. It makes sense this doesn't fluctuate, and it is more accurate than adding 1 in the denominator like our quick fix might have been in the earlier lesson. We will see that.

CASE & Aggregations

This one is pretty tricky. Try running the query yourself to make sure you understand what is happening. The next concept will give you some practice writing **CASE** statements on your own. In this section, we showed that getting the same information using a **WHERE** clause means only being able to get one set of data from the **CASE** at a time.

There are some advantages to separating data into separate columns like this depending on what you want to do, but often this level of separation might be easier to do in another programming language - rather than with SQL.

Example:

As a sales manager at Parch & Posey, classifying orders into general groups is a helpful exercise. But is much more useful if you can let it count up all the orders in each group. Aggregating based on these new will make it easier to report back to company leaders and take action. The easiest way to count all the members of a group is to create a column that group the way you want it to, then create another column to count by that group.

Here, we're using CASE to group orders into those with total quantity sold over 500, and those with 500 or less.

The screenshot shows a SQL editor interface with the following details:

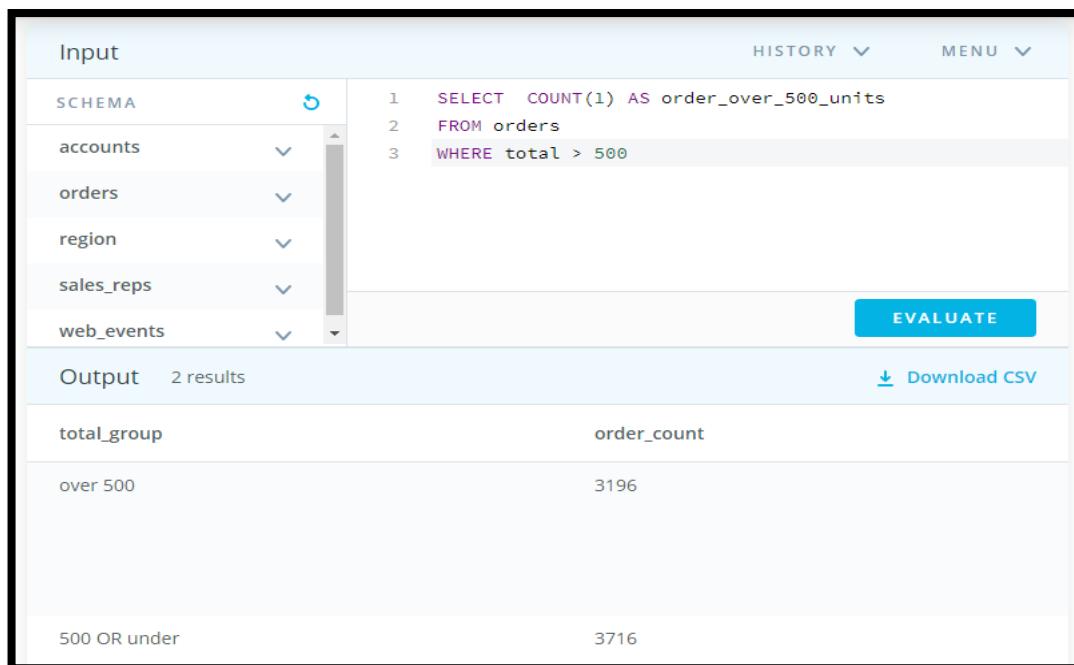
- Input:** A sidebar on the left lists tables: accounts, orders, region, sales_reps, and web_events. The 'orders' table is currently selected.
- Query Editor:** The main area contains the following SQL code:

```
1 SELECT CASE WHEN total > 500 THEN 'over 500' ELSE '500 OR under' END AS total_group, COUNT(*) AS order_count
2 FROM orders
3 GROUP BY 1
```
- Status:** Below the code, a green "Success!" message is displayed.
- EVALUATE:** A blue button labeled "EVALUATE" is located to the right of the status message.
- Output:** The results table shows two rows:

total_group	order_count
over 500	3196
500 OR under	3716
- Download CSV:** A link to download the results as a CSV file is located at the bottom right of the output area.

This query is an excellent place to use numbers instead of full column calculation in the GROUP BY clause, because repeating the CASE statement in the GROUP BY would make the query obviously long. Now, you might be thinking, why wouldn't I just use a WHERE clause to filter out rows I don't want to count?

You could do that, and it would look like this.



The screenshot shows a database query interface with the following details:

Input:

- SCHEMA dropdown menu with options: accounts, orders, region, sales_reps, web_events.
- Query code:

```
1 SELECT COUNT(1) AS order_over_500_units
2 FROM orders
3 WHERE total > 500
```
- EVALUATE button.

Output:

- 2 results
- Download CSV button.
- Data table:

total_group	order_count
over 500	3196
500 OR under	3716

Unfortunately, using the WHERE clause only allows you to count one condition at a time. This would be tedious if we had a number of different cases. We would need a separate query for each one.

Quiz: CASE

Questions: CASE

Use the **SQL** environment below to assist with answering the following questions.

1. We would like to understand 3 different levels of customers based on the amount associated with their purchases. The top branch includes anyone with a Lifetime Value (total sales of all orders) **greater than 200,000** usd. The second branch is between **200,000 and 100,000** usd. The lowest branch is anyone **under 100,000** usd. Provide a table that includes the **level** associated with each **account**. You should provide the **account name**, the **total sales of all orders** for the customer, and the **level**. Order with the top spending customers listed first.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1  SELECT a.name, SUM(total_amt_usd) total_spent,CASE
2      WHEN SUM(total_amt_usd) > 200000 THEN 'top' WHEN
3          SUM(total_amt_usd) > 100000 THEN 'middle' ELSE 'low'
4      END AS customer_level
5  FROM orders o
6  JOIN accounts a
7  ON o.account_id = a.id
8  GROUP BY a.name
9  ORDER BY 2 DESC;
```

Success! EVALUATE

Output 350 results Download CSV

name	total_spent	customer_level
EOG Resources	382873.30	top
Mosaic	345618.59	top
IBM	326819.48	top

2. We would now like to perform a similar calculation to the first, but we want to obtain the total amount spent by customers only in **2016** and **2017**. Keep the same **levels** as in the previous question. Order with the top spending customers listed first.

The screenshot shows a database interface with the following components:

- Input:** A sidebar on the left listing schemas: accounts, orders, region, sales_reps, and web_events. The sales_reps schema is currently selected.
- Query Editor:** The main area contains a SQL query:


```

1  SELECT a.name, SUM(total_amt_usd) total_spent, CASE
2      WHEN SUM(total_amt_usd) > 200000 THEN 'top'
3      WHEN SUM(total_amt_usd) > 100000 THEN 'middle'
4      ELSE 'low'
5  END AS customer_level
6  FROM orders o
7  JOIN accounts a
8  ON o.account_id = a.id
9  WHERE occurred_at > '2015-12-31'
10 GROUP BY 1
11 ORDER BY 2 DESC;
      
```
- Status:** Below the query editor, it says "Success!" and has a blue "EVALUATE" button.
- Output:** Below the status, it says "Output 322 results" and includes a "Download CSV" link.
- Results Table:** A table showing the output of the query:

name	total_spent	customer_level
Pacific Life	255319.18	top
Mosaic	172180.04	middle
CHS	163471.78	middle

3. We would like to identify top performing **sales reps**, which are sales reps associated with more than 200 orders. Create a table with the **sales rep name**, the total number of orders, and a column with **top** or **not** depending on if they have more than 200 orders. Place the top sales people first in your final table.

```

Input
SCHEMA
accounts
orders
region
sales_reps
web_events

1  SELECT s.name, COUNT(*) num_ords, CASE WHEN COUNT(*) > 200 THEN 'top' ELSE 'not' END AS sales_rep_level
2  FROM orders o
3  JOIN accounts a
4  ON o.account_id = a.id
5  JOIN sales_reps s
6  ON s.id = a.sales_rep_id
7  GROUP BY s.name
8  ORDER BY 2 DESC;

Success!
EVALUATE

Output 50 results
Download CSV

name          num_ords    sales_rep_level
Earlie Schleusner 335      top
Vernita Plump   299      top
Tia Amato       267      top

```

It is worth mentioning that this assumes each name is unique - which has been done a few times. We otherwise would want to break by the name and the id of the table.

4. The previous didn't account for the middle, nor the dollar amount associated with the sales. Management decides they want to see these characteristics represented as well. We would like to identify top performing **sales reps**, which are sales reps associated with more than **200** orders or more than **750000** in total sales. The **middle** group has any **rep** with more than **150** orders or **500000** in sales. Create a table with the **sales rep name**, the total number of orders, total sales across all orders, and a column with **top**, **middle**, or **low** depending on this criteria. Place the top sales people based on dollar amount of sales first in your final table. You might see a few upset sales people by this criteria!

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1  SELECT s.name, COUNT(*), SUM(o.total_amt_usd)
2    total_spent, CASE WHEN COUNT(*) > 200 OR
3      SUM(o.total_amt_usd) > 750000 THEN 'top' WHEN
4      COUNT(*) > 150 OR SUM(o.total_amt_usd) > 500000 THEN
5      'middle' ELSE 'low' END AS sales_rep_level
6
7  FROM orders o
8  JOIN accounts a
9  ON o.account_id = a.id
10 JOIN sales_reps s
11 ON s.id = a.sales_rep_id
12 GROUP BY s.name
13 ORDER BY 3 DESC;

```

Success! EVALUATE

Output 50 results Download CSV

name	count	total_spent	sales_rep_level
Earlie Schleusner	335	1098137.72	top
Tia Amato	267	1010690.60	top
Vernita Plump	299	934212.93	top
Georgianna Chisholm	256	886244.12	top
Arica Stoltzfus	186	810353.34	top
Dorotha Seawell	208	766935.04	top

You might see a few upset sales people by this criteria !

Recap

Each of the sections has been labeled to assist if you need to revisit a particular topic. Intentionally, the solutions for a particular section are actually not in the labeled section, because my hope is this will force you to practice if you have a question about a particular topic we covered.

You have now gained a ton of useful skills associated with **SQL**. The combination of **JOINS** and **Aggregations** are one of the reasons **SQL** is such a powerful tool. If there was a particular topic you struggled with, I suggest coming back and revisiting the questions with a fresh mind. The more you practice the better, but you also don't want to get stuck on the same problem for an extended period of time!

Introduction

What this lesson is about...

Up to this point you have learned a lot about working with data using SQL. **This lesson will focus on three topics:**

1. Subqueries
 2. Table Expressions
 3. Persistent Derived Tables
-

Both **subqueries** and **table expressions** are methods for being able to write a query that creates a table, and then write a query that interacts with this newly created table. Sometimes the question you are trying to answer doesn't have an answer when working directly with existing tables in database.

However, if we were able to create new tables from the existing tables, we know we could query these new tables to answer our question. This is where the queries of this lesson come to the rescue.

If you can't yet think of a question that might require such a query, don't worry because you are about to see a whole bunch of them!

Introduction to Subqueries

Whenever we need to use existing tables to create a new table that we then want to query again, this is an indication that we will need to use some sort of **subquery**. In the next couple of concepts, we will walk through an example together. Then you will get some practice tackling some additional problems on your own.

Some queries, also known as inner queries or nested queries, are a tool for performing operations in multiple steps.

For example, let's put our marketing manager hats back on. We'd like to know which channels send the most traffic per day on average to Parch & Posey.

In order to do that, we'll need to aggregate events by channel by day, then we need to take those and average them.

Write Your First Subquery

Your First Subquery

The first time you write a subquery it might seem really complex. Let's try breaking it down into its different parts.

We want to find the average number of events for each day for each channel. The first table will provide us the number of events for each day and channel, and then we will need to average these values together using a second query.

Example :

First, we'll start by querying the underlying table to make sure that date makes sense for what we're trying to do.

The screenshot shows a SQL query editor interface. At the top, there are buttons for 'Run', 'Limit 100', 'Format SQL', and 'View History...'. The query itself is:

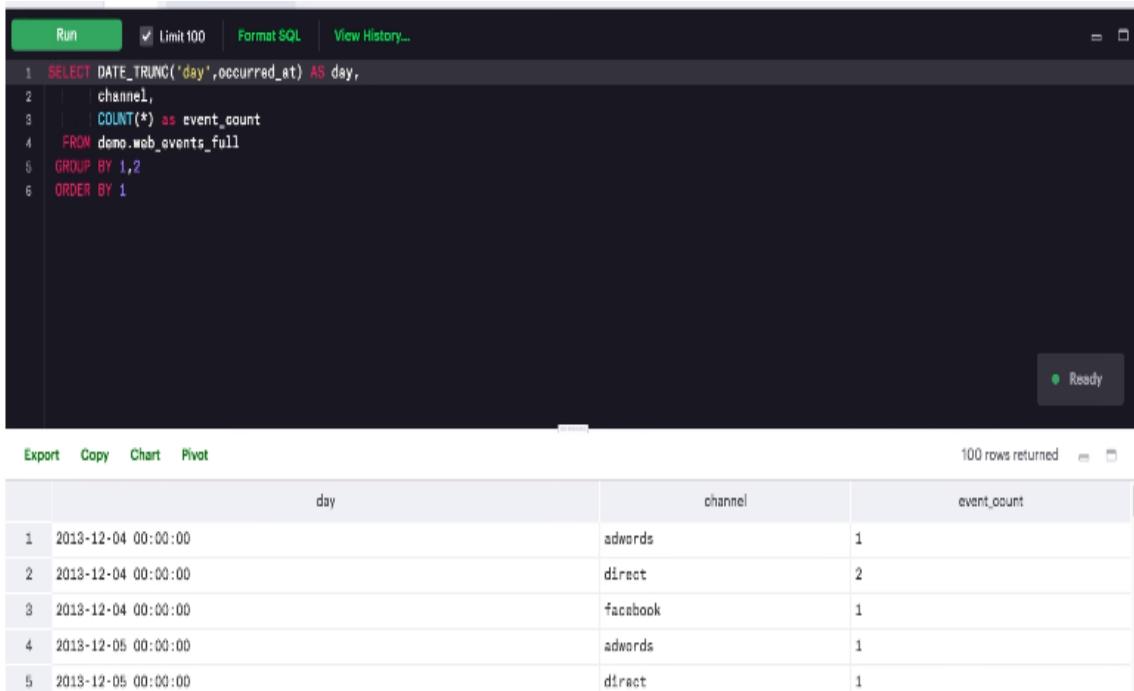
```
1 SELECT *
2 FROM demo.web_events_full
```

In the bottom right corner, a message indicates the query was successful: **Succeeded in 812ms**. Below the query results, there are buttons for 'Export', 'Copy', 'Chart', and 'Pivot'. The results table has columns: id, account_id, occurred_at, channel, and referrer_url. The data returned is:

	id	account_id	occurred_at	channel	referrer_url
1	4396	1001	2015-10-22 14:04:00	adwords	https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=order+paper
2	4399	1001	2016-01-01 15:45:00	adwords	https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=paper+company
3	4401	1001	2016-02-07 17:44:00	adwords	https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=parch+and+posey+paper
4	4410	1001	2016-06-22 13:48:00	adwords	https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=parch+and+posey+paper
5	4416	1001	2016-09-11 17:06:00	adwords	https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=parch+and+posey+paper

100 rows returned

Next, we'll count up all the events in each channel, in each day.



The screenshot shows a SQL query interface with the following details:

SQL Query:

```
1 SELECT DATE_TRUNC('day',occurred_at) AS day,
2      |   channel,
3      |   COUNT(*) AS event_count
4      FROM demo.web_events_full
5      GROUP BY 1,2
6      ORDER BY 1
```

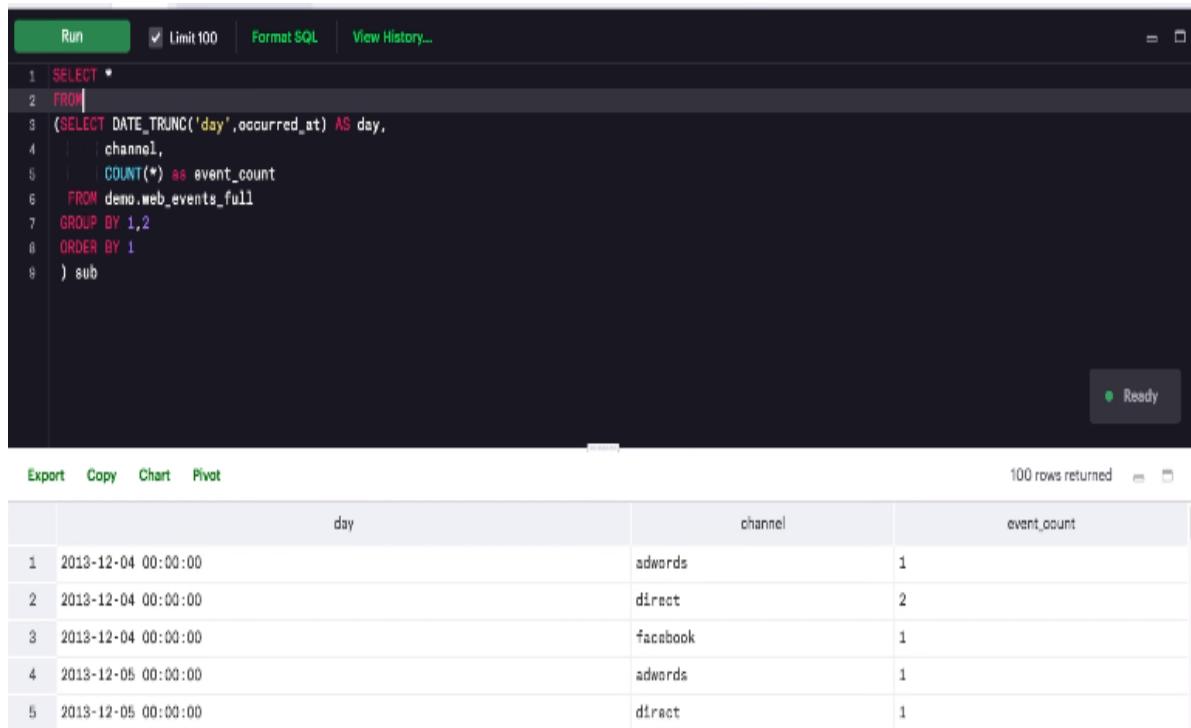
Results:

	day	channel	event_count
1	2013-12-04 00:00:00	adwords	1
2	2013-12-04 00:00:00	direct	2
3	2013-12-04 00:00:00	facebook	1
4	2013-12-05 00:00:00	adwords	1
5	2013-12-05 00:00:00	direct	1

Information at the bottom:

- Export Copy Chart Pivot
- 100 rows returned
- Ready

The last step is that we want to average across the events column we've created. In order to do that, we'll want to query against the results from this query. We can do that by wrapping the query in parentheses and using it in the FROM clause of the next query that we write.



The screenshot shows a SQL query interface with the following details:

SQL Query:

```
1 SELECT *
2 FROM (
3     SELECT DATE_TRUNC('day',occurred_at) AS day,
4           |   channel,
5           |   COUNT(*) AS event_count
6     FROM demo.web_events_full
7     GROUP BY 1,2
8     ORDER BY 1
9 ) sub
```

Results:

	day	channel	event_count
1	2013-12-04 00:00:00	adwords	1
2	2013-12-04 00:00:00	direct	2
3	2013-12-04 00:00:00	facebook	1
4	2013-12-05 00:00:00	adwords	1
5	2013-12-05 00:00:00	direct	1

Information at the bottom:

- Export Copy Chart Pivot
- 100 rows returned
- Ready

It is now a query within a query, also known as a subquery.

Subqueries are required to have aliases, which are added after the parentheses, the same way you would add an alias to a table. Here, we're just selecting all of the data from the subquery.

Let's go the last mile and average for each channel. Since the subquery acts like one table in the FROM clause, we'll put a GROUP BY clause after the subquery.

The screenshot shows a SQL editor interface with the following details:

- Toolbar:** Run, Limit 100, Format SQL, View History...
- Query Editor:** Contains the following SQL code:

```
1 SELECT channel,
2       AVG(event_count) AS avg_event_count
3   FROM
4   (SELECT DATE_TRUNC('day', occurred_at) AS day,
5        channel,
6        COUNT(*) AS event_count
7     FROM demo.web_events_full
8    GROUP BY 1,2
9   ORDER BY 1
10 ) sub
11   GROUP BY 1
12  ORDER BY 2 DESC
```
- Status Bar:** ✓ Succeeded in 1s
- Bottom Buttons:** Export, Copy, Chart, Pivot
- Bottom Right:** 6 rows returned
- Table Results:** A table with two columns: "channel" and "avg_event_count". The data is as follows:

	channel	avg_event_count
1	direct	4.096487985212589
2	organic	1.6672504378283712
3	facebook	1.5983471074380164
4	adwords	1.5701906412478337
5	twitter	1.3166666666666667

Since we're now reordering based on this new aggregation, we no longer need the ORDER BY statement in the subquery. So let's take that out to keep things clean.

The screenshot shows a SQL editor interface with the following details:

- Query:**

```

1 SELECT channel,
2       AVG(event_count) AS avg_event_count
3 FROM
4 (SELECT DATE_TRUNC('day',occurred_at) AS day,
5        channel,
6        COUNT(*) AS event_count
7   FROM demo.web_events_full
8  GROUP BY 1,2
9 ) sub
10 GROUP BY 1
11 ORDER BY 2 DESC

```
- Status:** ✓ Succeeded in 890ms
- Results:** A table with 6 rows returned, showing the average event count for four channels.

	channel	avg_event_count
1	direct	4.898407985212589
2	organic	1.8672504378283712
3	facebook	1.5983471074380164
4	adwords	1.5701906412478337

Just to make it super clear, let's break down how this new query runs.

First, your inner query will run. It might sound like a no-brainer, but it's important. Your inner query must actually run on its own as the database will treat it as an independent query. Once your inner query is complete, the rest of the query also known as the outer query will run across the result set created by the inner query.

A nice feature that mode and many other SQL editors share is the ability to highlight a portion of the query and run only that portion. This is especially helpful when making changes to an inner query. You can make the change, than quickly check the inner queries output to make sure it looks correct before running the outer query again.

Quiz: Write Your First Subquery

Quiz 1

On which day-channel pair did the most events occur. (Mark all that are true)

- January 1, 2017; direct

- December 31, 2016; facebook

- November 3, 2016; direct

- December 21, 2016; direct

Quiz 2

Mark all of the below that are true regarding writing your subquery.

- The original query goes in the **FROM** statement.

- The original query goes in the **SELECT** statement.

- An * is used in the **SELECT** statement to pull all of the data from the original query.

- You **MUST** use an alias for the table you nest within the outer query.

QUESTION 3 OF 3

Match each channel to its corresponding average number of events per day.

Submit to check your answer choices!

CHANNEL	AVERAGE NUMBER OF EVENTS/DAY
Direct	4.90
Facebook	1.60
Organic	1.67
Twitter	1.32

You try solving this yourself.

Tasks to complete.

Task List

1. Use the test environment below to find the number of events that occur for each day for each channel.

The screenshot shows a data analysis interface with the following components:

- Input:** A sidebar on the left lists tables: accounts, orders, region, sales_reps, and web_events. The "web_events" table is currently selected.
- Query Editor:** The main area contains a SQL query:

```
1 SELECT DATE_TRUNC('day', occurred_at) AS day,
2   channel, COUNT(*) AS events
3   FROM web_events
4   GROUP BY 1,2
5   ORDER BY 3 DESC;
```
- Status:** Below the query editor, it says "Success!"
- EVALUATE:** A blue button to run the query.
- Output:** Shows "3564 results".
- Download CSV:** A link to download the results.
- Results Table:** A table with columns: day, channel, and events. The data is:

day	channel	events
2017-01-01T00:00:00.000Z	direct	21
2016-12-21T00:00:00.000Z	direct	21
2016-12-31T00:00:00.000Z	direct	19

2. Now create a subquery that simply provides all of the data from your first query.

The screenshot shows a database query interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events
- Query:**

```
1 SELECT *
2 FROM (SELECT DATE_TRUNC('day',occurred_at) AS day,
3 channel, COUNT(*) as events
4 FROM web_events
5 GROUP BY 1,2
6 ORDER BY 3 DESC) sub;
```

- Status:** Success!
- EVALUATE** button
- Output:** 3564 results
- Download CSV** link
- Table Data:**

day	channel	events
2017-01-01T00:00:00.000Z	direct	21
2016-12-21T00:00:00.000Z	direct	21
2016-12-31T00:00:00.000Z	direct	19

3. Now find the average number of events for each channel. Since you broke out by day earlier, this is giving you an average per day.

The screenshot shows a database query interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events
- Query:**

```
1 SELECT channel, AVG(events) AS average_events
2 FROM (SELECT DATE_TRUNC('day',occurred_at) AS day,
3 channel, COUNT(*) as events
4 FROM web_events
5 GROUP BY 1,2) sub
6 GROUP BY channel
7 ORDER BY 2 DESC;
```

- Status:** Success!
- EVALUATE** button
- Output:** 6 results
- Download CSV** link
- Table Data:**

channel	average_events
direct	4.8964879852125693
organic	1.6672504378283713
facebook	1.5983471074380165

Subquery Formatting

Subquery Formatting

When writing **Subqueries**, it is easy for your query to look incredibly complex. In order to assist your reader, which is often just yourself at a future date, formatting SQL will help with understanding your code.

The important thing to remember when using subqueries is to provide some way for the reader to easily determine which parts of the query will be executed together. Most people do this by indenting the subquery in some way - you saw this with the solution blocks in the previous concept.

The examples in this class are indented quite far—all the way to the parentheses. This isn't practical if you nest many subqueries, but in general, be thinking about how to write your queries in a readable way. Examples of the same query written multiple different ways is provided below. You will see that some are much easier to read than others.

Badly Formatted Queries

Though these poorly formatted examples will execute the same way as the well formatted examples, they just aren't very friendly for understanding what is happening!

Here is the first, where it is impossible to decipher what is going on:

```
SELECT * FROM (SELECT DATE_TRUNC('day',occurred_at) AS day, channel,  
COUNT(*) as events FROM web_events GROUP BY 1,2 ORDER BY 3 DESC) sub;
```

This second version, which includes some helpful line breaks, is easier to read than that previous version, but it is still not as easy to read as the queries in the [Well Formatted Query](#) section.

```
SELECT *
FROM (
  SELECT DATE_TRUNC('day',occurred_at) AS day,
  channel, COUNT(*) as events
  FROM web_events
  GROUP BY 1,2
  ORDER BY 3 DESC) sub;
```

Well Formatted Query

Now for a well formatted example, you can see the table we are pulling from much easier than in the previous queries.

```
SELECT *
FROM (SELECT DATE_TRUNC('day',occurred_at) AS day,
  channel, COUNT(*) as events
  FROM web_events
  GROUP BY 1,2
  ORDER BY 3 DESC) sub;
```

Additionally, if we have a [GROUP BY](#), [ORDER BY](#), [WHERE](#), [HAVING](#), or any other statement following our subquery, we would then indent it at the same level as our outer query.

The query below is similar to the above, but it is applying additional statements to the outer query, so you can see there are [GROUP BY](#) and [ORDER BY](#) statements used on the output are not tabbed. The inner query [GROUP BY](#) and [ORDER BY](#) statements are indented to match the inner table.

```
SELECT *  
FROM (SELECT DATE_TRUNC('day',occurred_at) AS day,  
       channel, COUNT(*) as events  
      FROM web_events  
     GROUP BY 1,2  
    ORDER BY 3 DESC) sub  
   GROUP BY channel  
 ORDER BY 2 DESC;
```

These final two queries are so much easier to read!

More On Subqueries

Subqueries Part II

In the first subquery you wrote, you created a table that you could then query again in the **FROM** statement. However, if you are only returning a single value, you might use that value in a logical statement like **WHERE**, **HAVING**, or even **SELECT**- the value could be nested within a **CASE** statement.

On the next concept, we will work through this example, and then you will get some practice on answering some questions on your own.

Expert Tip

Note that you should not include an alias when you write a subquery in a conditional statement. This is because the subquery is treated as an individual value (or set of values in the **IN** case) rather than as a table.

Also, notice the query here compared a single value. If we returned an entire column **IN** would need to be used to perform a logical argument. If we are returning an entire table, then we must use an **ALIAS** for the table, and perform additional logic on the entire table.

Subqueries can be used in several places within a query. It can really be used anywhere you might use a table name or even a column name or an individual value. They're especially useful in conditional logic, in conjunction with where or clauses, or in the when portion of a case statement .

For example, you might want to return only orders that occurred in the same month as Parch & Posey's first order ever. To get the date of the first order, you can write a subquery with a min function.

Input

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT MIN(occurred_at) AS min
2 FROM orders
```

Success!

EVALUATE

Output 1 results

min

2013-12-04T04:22:44.000Z

Let's add a date_trunc function to get the month.

Run Limit 100 Format SQL View History...

```
1 SELECT DATE_TRUNC('month',MIN(occurred_at)) AS min_month
2 FROM demo.orders
```

Ready

Export Copy Chart Pivot 1 rows returned

	min_month
1	2013-12-01 00:00:00

Finally, let's write an outer query that uses this to filter the orders table and sorts by the occurred at column.

The screenshot shows a SQL query being run in a database environment. The query is:

```

1 SELECT *
2 FROM demo.orders
3 WHERE DATE_TRUNC('month',occurred_at) =
4 (SELECT DATE_TRUNC('month',MIN(occurred_at)) AS min_month
5 FROM demo.orders )
6 ORDER BY occurred_at

```

The result of the query is displayed below, showing 99 rows returned. The data includes columns such as id, account_id, occurred_at, standard_qty, gloss_qty, poster_qty, total, standard_amt_usd, gloss_amt_usd, poster_amt_usd, and total_amt_usd. All rows show occurrences in December 2013.

Export Copy Chart Pivot

99 rows returned

	id	account_id	occurred_at	standard_qty	gloss_qty	poster_qty	total	standard_amt_usd	gloss_amt_usd	poster_amt_usd	total_amt_usd
1	5786	2961	2013-12-04 04:22:44	0	48	33	81	0	359.52	287.98	627.48
2	2415	2961	2013-12-04 04:45:54	490	15	11	516	2445.1	112.35	89.32	2648.77
3	4108	4311	2013-12-04 04:53:25	528	10	0	538	2634.72	74.9	0	2709.62
4	4489	1281	2013-12-05 20:29:16	0	37	0	37	0	277.13	0	277.13

You can see that all of the results took place in December 2013, the same month as the first order. This query works because the result of the subquery is only one cell. Most conditional logic will work with subqueries containing one-cell results. But IN is the only type of conditional logic that will work when the inner query contains multiple results.

Quiz: More On Subqueries

Use the task list below to work through the previous example.

Tasks to complete:

Use **DATE_TRUNC** to pull **month** level information about the first order ever placed in the **orders** table.

The screenshot shows a database query editor interface. The top bar includes 'HISTORY' and 'MENU' dropdowns. The main area has a 'Input' tab where a schema dropdown is set to 'accounts'. Below it is a code editor with two lines of SQL:

```
1 SELECT DATE_TRUNC('month', MIN(occurred_at))  
2 FROM orders
```

A green 'Success!' message is displayed next to the code editor. A large blue 'EVALUATE' button is located to the right of the message. The bottom section is labeled 'Output' and shows one result row:

date_trunc
2013-12-01T00:00:00.000Z

Use the result of the previous query to find only the orders that took place in the same month and year as the first order, and then pull the average for each type of paper **qty** in this month.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT AVG(standard_qty) avg_std, AVG(gloss_qty)
      avg_gls, AVG(poster_qty) avg_pst
2 FROM orders
3 WHERE DATE_TRUNC('month', occurred_at) = (SELECT
      DATE_TRUNC('month', MIN(occurred_at)) FROM orders);
4 SELECT SUM(total_amt_usd)
5 FROM orders
6 WHERE DATE_TRUNC('month', occurred_at) = (SELECT
      DATE_TRUNC('month', MIN(occurred_at)) FROM orders);
```

Success!

EVALUATE

Output 1 results

sum

377331.00

QUESTION 1 OF 2

What was the month/year combo for the first order placed?

October 2011

October 2012

December 2013

November 2013

QUESTION 2 OF 2

Match each value to the corresponding description.

DESCRIPTION

VALUE

The average amount of standard paper sold
on the first month that any order was placed in
the **orders** table (in terms of quantity).

268

The average amount of gloss paper sold on
the first month that any order was placed in
the **orders** table (in terms of quantity).

209

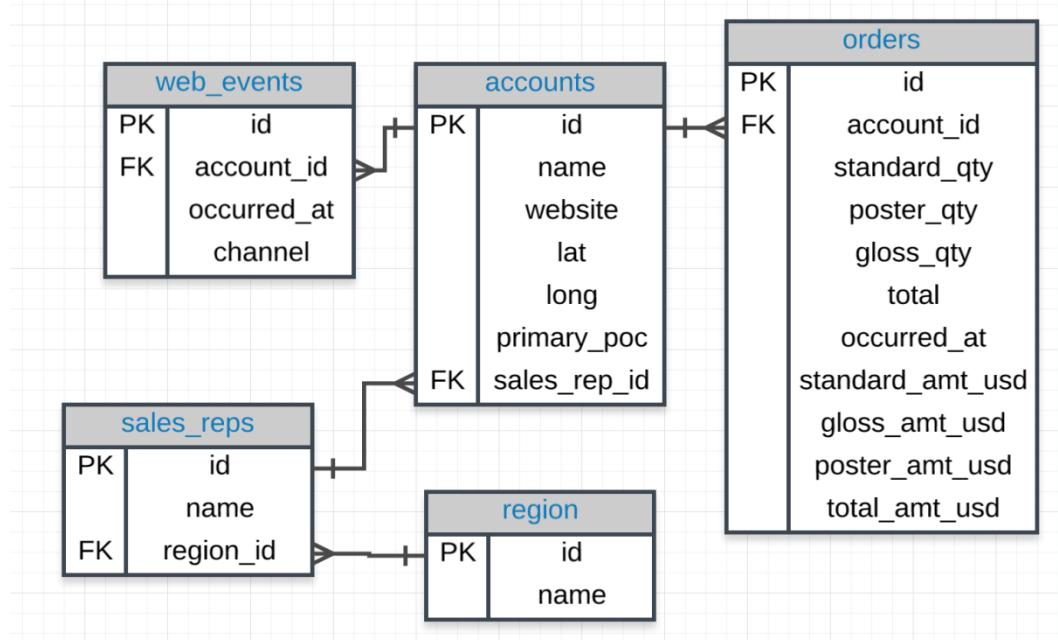
The average amount of poster paper sold on
the first month that any order was placed in
the **orders** table (in terms of quantity).

112

The total amount spent on all orders on the
first month that any order was placed in the
orders table (in terms of usd).

377331

Quiz: Subquery Mania



More Subqueries Quizzes

Above is the ERD for the database again - it might come in handy as you tackle the quizzes below. You should write your solution as a subquery, not by finding one solution and copying the output. The importance of this is that it allows your query to be dynamic in answering the question - even if the data changes, you still arrive at the right answer.

1. Provide the name of the **sales_rep** in each **region** with the largest amount of **total_amt_usd** sales.

First, I wanted to find the **total_amt_usd** totals associated with each **sales rep**, and I also wanted the region in which they were located. The query below provided this information.

Input

SCHEMA	
accounts	v
orders	v
region	v
sales_reps	v
web_events	v

```
1  SELECT s.name rep_name, r.name region_name,
2      SUM(o.total_amt_usd) total_amt
3  FROM sales_reps s
4  JOIN accounts a
5  ON a.sales_rep_id = s.id
6  JOIN orders o
7  ON o.account_id = a.id
8  JOIN region r
9  ON r.id = s.region_id
10 GROUP BY 1,2
11 ORDER BY 3 DESC;
```

Success! EVALUATE

Output 50 results Download CSV

rep_name	region_name	total_amt
Earlie Schleusner	Southeast	1098137.72
Tia Amato	Northeast	1010690.60
Vernita Plump	Southeast	934212.93
Georgianna Chisholm	West	886244.12
Arica Stoltzfus	West	810353.34
Dorotha Seawell	Southeast	766935.04

Next, I pulled the max for each region, and then we can use this to pull those rows in our final result.

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1  SELECT s.name rep_name, r.name region_name,
2      SUM(o.total_amt_usd) total_amt
3  FROM sales_reps s
4  JOIN accounts a
5  ON a.sales_rep_id = s.id
6  JOIN orders o
7  ON o.account_id = a.id
8  JOIN region r
9  ON r.id = s.region_id
10 GROUP BY 1,2
11 ORDER BY 3 DESC;

```

Success!

EVALUATE

Output 50 results

Download CSV

rep_name	region_name	total_amt
Earlie Schleusner	Southeast	1098137.72
Tia Amato	Northeast	1010690.60
Vernita Plump	Southeast	934212.93
Georgianna Chisholm	West	886244.12
Arica Stoltzfus	West	810353.34
Dorotha Seawell	Southeast	766935.04
Nelle Meaux	Southeast	749076.16
Sibyl Lauria	Northeast	722084.27

^MENU

Essentially, this is a **JOIN** of these two tables, where the region and amount match.

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
orders	▼
region	▼
region	▼
sales_reps	▼
web_events	▼

```

1  SELECT t3.rep_name, t3.region_name, t3.total_amt
2  FROM(SELECT region_name, MAX(total_amt) total_amt
3          FROM(SELECT s.name rep_name, r.name region_name,
4                 SUM(o.total_amt_usd) total_amt
5             FROM sales_reps s
6             JOIN accounts a
7                 ON a.sales_rep_id = s.id
8             JOIN orders o
9                 ON o.account_id = a.id
10            JOIN region r
11               ON r.id = s.region_id
12           GROUP BY 1, 2) t1
13           GROUP BY 1) t2
14   JOIN (SELECT s.name rep_name, r.name region_name,
15          SUM(o.total_amt_usd) total_amt
16      FROM sales_reps s
17      JOIN accounts a
18         ON a.sales_rep_id = s.id
19      JOIN orders o
20         ON o.account_id = a.id
21         JOIN region r
22            ON r.id = s.region_id
23           GROUP BY 1,2

```

EVALUATE

Output 4 results [Download CSV](#)

rep_name	region_name	total_amt
Earlie Schleusner	Southeast	1098137.72
Tia Amato	Northeast	1010690.60
Georgianna Chisholm	West	886244.12
Charles Bidwell	Midwest	675637.19

2. For the region with the largest (sum) of sales **total_amt_usd**, how many **total** (count) orders were placed?

The first query I wrote was to pull the **total_amt_usd** for each **region**.

The screenshot shows a database query editor interface. At the top, there's a header with "Input", "HISTORY ▾", and "MENU ▾". Below the header, on the left, is a sidebar titled "SCHEMA" with a circular refresh icon. It lists several tables: "accounts" (selected), "orders", "region" (highlighted with a blue border), "sales_reps", and "web_events". On the right side of the interface, the query code is displayed:

```

1  SELECT r.name region_name, SUM(o.total_amt_usd)
2    total_amt
3  FROM sales_reps s
4  JOIN accounts a
5  ON a.sales_rep_id = s.id
6  JOIN orders o
7  ON o.account_id = a.id
8  JOIN region r
9  ON r.id = s.region_id
   GROUP BY r.name;

```

Below the code, a green "Success!" message is shown next to a blue "EVALUATE" button.

Under the "Output" section, it says "4 results". The results are presented in a table:

region_name	total_amt
Midwest	3013486.51
Southeast	6458497.00
Northeast	7744405.36
West	5925122.96

Then we just want the region with the max amount from this table. There are two ways I considered getting this amount. One was to pull the max using a subquery. Another way is to order descending and just pull the top value.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT MAX(total_amt)
2 FROM (SELECT r.name region_name, SUM(o.total_amt_usd)
3        total_amt
4     FROM sales_reps s
5   JOIN accounts a
6      ON a.sales_rep_id = s.id
7   JOIN orders o
8      ON o.account_id = a.id
9   JOIN region r
10     ON r.id = s.region_id
11 GROUP BY r.name) sub;
```

Success!

EVALUATE

Output 1 results

max

7744405.36

Finally, we want to pull the total orders for the region with this amount:

Input

HISTORY ▾

MENU ▾

SCHEMA

accounts

orders

sales_reps

web_events

```
1  SELECT r.name, COUNT(o.total) total_orders
2  FROM sales_reps s
3  JOIN accounts a
4  ON a.sales_rep_id = s.id
5  JOIN orders o
6  ON o.account_id = a.id
7  JOIN region r
8  ON r.id = s.region_id
9  GROUP BY r.name
10 HAVING SUM(o.total_amt_usd) = (
11     SELECT MAX(total_amt)
12     FROM (SELECT r.name region_name,
13                 SUM(o.total_amt_usd) total_amt
14             FROM sales_reps s
15             JOIN accounts a
16             ON a.sales_rep_id = s.id
17             JOIN orders o
18             ON o.account_id = a.id
19             JOIN region r
20             ON r.id = s.region_id
21         GROUP BY r.name) sub);
```

EVALUATE

Output 1 results

name

total_orders

Northeast

2357

3. For the **name** of the account that purchased the most (in total over their lifetime as a customer) **standard_qty**, **how many accounts** still had more in **total** purchases?

First, we want to find the account that had the most **standard_qty** paper. **The query here pulls that account, as well as the total amount:**

The screenshot shows a database query editor interface. The top navigation bar includes 'Input', 'HISTORY ▾', and 'MENU ▾'. On the left, there's a 'SCHEMA' dropdown with options: 'accounts', 'orders', 'region', 'sales_reps', and 'web_events'. The main area displays a SQL query:

```
1 SELECT a.name account_name, SUM(o.standard_qty)
  2   total_std, SUM(o.total) total
  3   FROM accounts a
  4   JOIN orders o
  5   ON o.account_id = a.id
  6   GROUP BY 1
  7   ORDER BY 2 DESC
  8   LIMIT 1;
```

A green 'Success!' message is displayed below the query. To the right is a blue 'EVALUATE' button. The bottom section is labeled 'Output 1 results' and shows a table with three columns: 'account_name', 'total_std', and 'total'. The single result row is:

account_name	total_std	total
Core-Mark Holding	41617	44750

Now, I want to use this to pull all the accounts with more total sales:

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT a.name
2 FROM orders o
3 JOIN accounts a
4 ON a.id = o.account_id
5 GROUP BY 1
6 HAVING SUM(o.total) > (SELECT total
7 FROM (SELECT a.name act_name, SUM(o.standard_qty)
8 tot_std, SUM(o.total) total
9 FROM accounts a
10 JOIN orders o
11 ON o.account_id = a.id
12 GROUP BY 1
13 ORDER BY 2 DESC
14 LIMIT 1) sub);
```

Success!

EVALUATE

Output 3 results

name
Mosaic
EOG Resources
IBM

This is now a list of all the accounts with more total orders. We can get the count with just another simple subquery.

Input

HISTORY ▾

MENU ▾

SCHEMA

accounts

orders

region

sales_reps

web_events

```
1  SELECT COUNT(*)
2  FROM (SELECT a.name
3          FROM orders o
4          JOIN accounts a
5            ON a.id = o.account_id
6        GROUP BY 1
7        HAVING SUM(o.total) > (SELECT total
8                               FROM (SELECT a.name act_name,
9                                     SUM(o.standard_qty) tot_std, SUM(o.total) total
10                                 FROM accounts a
11                                JOIN orders o
12                                  ON o.account_id = a.id
13                                GROUP BY 1
14                                ORDER BY 2 DESC
15                                LIMIT 1) inner_tab)
15      ) counter_tab;
```

Success!

EVALUATE

Output 1 results

count

3

4. For the customer that spent the most (in total over their lifetime as a customer) **total_amt_usd**, how many **web_events** did they have for each channel?

Here, we first want to pull the customer with the most spent in lifetime value.

The screenshot shows a database query interface with the following details:

Input:

- SCHEMA dropdown: 5 schemas listed: accounts, orders, region, sales_reps, web_events.
- Query code:

```
1 SELECT a.id, a.name, SUM(o.total_amt_usd) tot_spent
2 FROM orders o
3 JOIN accounts a
4 ON a.id = o.account_id
5 GROUP BY a.id, a.name
6 ORDER BY 3 DESC
7 LIMIT 1;
```
- Status: Success!
- EVALUATE button.

Output: 1 results

id	name	tot_spent
4211	EOG Resources	382873.30

Now, we want to look at the number of events on each channel this company had, which we can match with just the **id**.

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1  SELECT a.name, w.channel, COUNT(*)
2  FROM accounts a
3  JOIN web_events w
4  ON a.id = w.account_id AND a.id = (SELECT id
5                      FROM (SELECT a.id, a.name,
6                            SUM(o.total_amt_usd) tot_spent
7                      FROM orders o
8                      JOIN accounts a
9                         ON a.id = o.account_id
10                     GROUP BY a.id, a.name
11                     ORDER BY 3 DESC
12                     LIMIT 1) inner_table)
13  GROUP BY 1, 2
14  ORDER BY 3 DESC;

```

Success! EVALUATE

Output 6 results

name	channel	count
EOG Resources	direct	44
EOG Resources	organic	13
EOG Resources	adwords	12
EOG Resources	facebook	11
EOG Resources	twitter	5
EOG Resources	banner	4

I added an **ORDER BY** for no real reason, and the account name to assure I was only pulling from one account.

5. What is the lifetime average amount spent in terms of **total_amt_usd** for the top 10 total spending **accounts**?

First, we just want to find the top 10 accounts in terms of highest **total_amt_usd**.

The screenshot shows a database query interface with the following details:

Input:

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

Query code:

```
1 SELECT a.id, a.name, SUM(o.total_amt_usd) tot_spent
2 FROM orders o
3 JOIN accounts a
4 ON a.id = o.account_id
5 GROUP BY a.id, a.name
6 ORDER BY 3 DESC
7 LIMIT 10;
```

Status: Success!

EVALUATE

Output: 10 results

id	name	tot_spent
4211	EOG Resources	382873.30
4151	Mosaic	345618.59
1301	IBM	326819.48
1871	General Dynamics	300694.79
4111	Republic Services	293861.14

Now, we just want the average of these 10 amounts.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT AVG(tot_spent)
2 FROM (SELECT a.id, a.name, SUM(o.total_amt_usd)
        tot_spent
3         FROM orders o
4         JOIN accounts a
5     ON a.id = o.account_id
6     GROUP BY a.id, a.name
7     ORDER BY 3 DESC
8     LIMIT 10) temp;
```

Success!

EVALUATE

Output 1 results

avg
304846.969000000000

6. What is the lifetime average amount spent in terms of **total_amt_usd** for only the companies that spent more than the average of all orders.

First, we want to pull the average of all accounts in terms of **total_amt_usd**:

The screenshot shows a database query editor interface. On the left, there is a sidebar titled "Input" with a "SCHEMA" dropdown menu containing five entries: "accounts", "orders", "region", "sales_reps", and "web_events". Below the schema list is a status message "Success!" and a blue "EVALUATE" button. The main area is titled "Output" and shows 1 result. The result is a single row with one column labeled "avg_all" containing the value "3348.0196513310185185".

```
1 SELECT AVG(o.total_amt_usd) avg_all
2 FROM orders o
3 JOIN accounts a
4 ON a.id = o.account_id;
```

Success!

EVALUATE

Output 1 results

avg_all

3348.0196513310185185

Then, we want to only pull the accounts with more than this average amount.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT o.account_id, AVG(o.total_amt_usd)
2 FROM orders o
3 GROUP BY 1
4 HAVING AVG(o.total_amt_usd) > (SELECT
5          AVG(o.total_amt_usd) avg_all
6          FROM orders o
7          JOIN accounts a
8          ON a.id =
9
Success!
```

EVALUATE

Output 169 results

account_id	avg
2651	5106.0793750000000000
2941	5389.8850000000000000
1501	3993.2714285714285714
1351	4440.7220000000000000

Finally, we just want the average of these values.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT AVG(avg_amt)
2 FROM (SELECT o.account_id, AVG(o.total_amt_usd)
3        FROM orders o
4        GROUP BY 1
5        HAVING AVG(o.total_amt_usd) > (SELECT
6              AVG(o.total_amt_usd) avg_all
7              FROM orders o
8              JOIN accounts a
9              ON a.id =
10             o.account_id)) temp_table;
11
Success!
```

EVALUATE

Output 1 results

avg
4721.1397439971747168

WITH

The **WITH** statement is often called a **Common Table Expression** or **CTE**. Though these expressions serve the exact same purpose as subqueries, they are more common in practice, as they tend to be cleaner for a future reader to follow the logic.

In the next concept, we will walk through this example a bit more slowly to make sure you have all the similarities between subqueries and these expressions down for you to use in practice!

If you have a common table expression or subquery that takes a really long time to run, you might want to run it as completely separate query and then write it back into the database at its own table. Then, you can simply query the new table as you would any other, to finish the thing you're trying to calculate.

The big benefit to this approach is that you can improve the speed at which you explore.

Let's say you've got a subquery that takes an hour to write, and then you've got an outer query that's pretty fast. If you run both of them every time you want to make a small tweak to the outer query, you're going to get really frustrated.

Instead, you can run that inner query once and write it to a table, then iterating on the outer query will be quick and easy. We'll dig into performance in more depth in later lesson.

For now, we'll focus on writing your subqueries and common table expressions into the database as new tables.

WITH vs. Subquery

Your First WITH (CTE)

The same question as you saw in [your first subquery](#) is provided here along with the solution.

QUESTION: You need to find the average number of events for each channel per day.

SOLUTION:

The screenshot shows a SQL query editor interface. On the left, there's a sidebar titled "Input" with a "SCHEMA" dropdown menu containing options: accounts, orders, region, sales_reps, and web_events. The main area contains a SQL query:

```
1 SELECT channel, AVG(events) AS average_events
2 FROM (SELECT DATE_TRUNC('day', occurred_at) AS day,
3             channel, COUNT(*) as events
4      FROM web_events
5     GROUP BY 1,2) sub
6  GROUP BY channel
7 ORDER BY 2 DESC;
```

The status bar at the bottom of the editor says "Success!" and has a blue "EVALUATE" button. Below the editor, there's an "Output" section showing the results of the query:

channel	average_events
direct	4.8964879852125693
organic	1.6672504378283713
facebook	1.5983471074380165

There is also a "Download CSV" link next to the output table.

Let's try this again using a **WITH** statement.

Notice, you can pull the inner query:

The screenshot shows a data analysis tool interface. In the top left, there's a sidebar titled "Input" with a "SCHEMA" dropdown showing "accounts", "orders", "region", "sales_reps", and "web_events". The main area contains a query editor with the following code:

```
1  SELECT DATE_TRUNC('day',occurred_at) AS day,
2          channel, COUNT(*) as events
3  FROM web_events
4  GROUP BY 1,2
```

The status bar at the bottom of the editor says "Success!". To the right of the editor is a blue "EVALUATE" button. Below the editor is a section titled "Output" showing "3564 results". A "Download CSV" button is next to it. The output table has three columns: "day", "channel", and "events". The data is as follows:

day	channel	events
2015-11-04T00:00:00.000Z	banner	1
2013-12-11T00:00:00.000Z	organic	1
2016-01-06T00:00:00.000Z	direct	8
2016-05-07T00:00:00.000Z	facebook	5

This is the part we put in the **WITH** statement. Notice, we are aliasing the table as **events** below:

```
WITH events AS (
    SELECT DATE_TRUNC('day',occurred_at) AS day,
          channel, COUNT(*) as events
    FROM web_events
    GROUP BY 1,2)
```

Now, we can use this newly created **events** table as if it is any other table in our database:

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1 WITH events AS (SELECT DATE_TRUNC('day',occurred_at)
2           AS day, channel, COUNT(*) as events
3           FROM web_events
4           GROUP BY 1,2)
5   SELECT channel, AVG(events) AS average_events
6   FROM events
7   GROUP BY channel
8   ORDER BY 2 DESC;

```

Success!

EVALUATE

Output 6 results

[Download CSV](#)

channel	average_events
direct	4.8964879852125693
organic	1.6672504378283713
facebook	1.5983471074380165

For the above example, we don't need anymore than the one additional table, but imagine we needed to create a second table to pull from. **We can create an additional table to pull from in the following way:**

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1 WITH table1 AS (
2   SELECT *
3   FROM web_events),
4   table2 AS (
5   SELECT *
6   FROM accounts)
7   SELECT *
8   FROM table1
9   JOIN table2
10  ON table1.account_id = table2.id;

```

EVALUATE

Output 9073 results

[Download CSV](#)

id	account_id	occurred_at	channel	name
1001	1001	2015-10-06T17:13:58.000Z	direct	Walmart
1001	1001	2015-11-05T03:08:26.000Z	direct	Walmart
1001	1001	2015-12-04T03:57:24.000Z	direct	Walmart

You can add more and more tables using the **WITH** statement in the same way. The quiz at the bottom will assure you are catching all of the necessary components of these new queries.

Quiz: *WITH vs. Subquery*

QUIZ QUESTION

Select all of the below that are true regarding **WITH** statements.

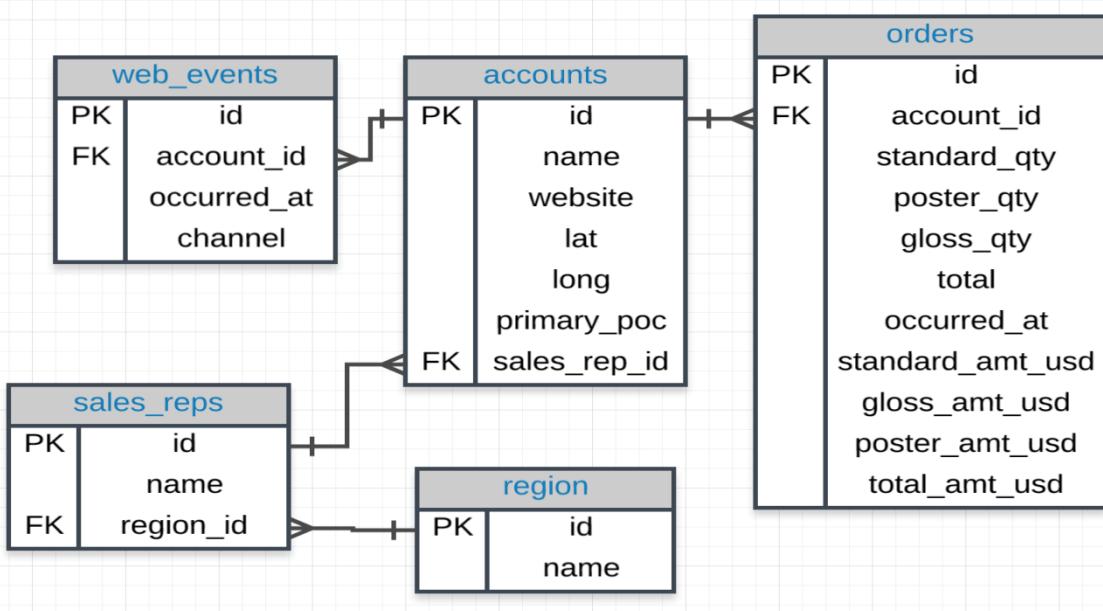
- When creating multiple tables using **WITH**, you add a comma after every table leading to your final query.

- When creating multiple tables using **WITH**, you add a comma after every table except the last table leading to your final query.

- The new table name is always aliased using `table_name AS`, which is followed by your query nested between parentheses.

- You begin each new table using a **WITH** statement.

Quiz: *WITH*



Essentially a **WITH** statement performs the same task as a **Subquery**. Therefore, you can write any of the queries we worked with in the "Subquery Mania" using a **WITH**. That's what you'll do here. Try to perform each of the earlier queries again, but using a **WITH** instead of a subquery.

Above is the ERD for the database again - it might come in handy as you tackle the quizzes below. You should write your solution as using a **WITH** statement, not by finding one solution and copying the output. The importance of this is that it allows your query to be dynamic in answering the question - even if the data changes, you still arrive at the right answer.

1. Provide the **name** of the **sales_rep** in each **region** with the largest amount of **total_amt_usd** sales.

Input

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 WITH t1 AS (
2     SELECT s.name rep_name, r.name region_name,
3            SUM(o.total_amt_usd) total_amt
4     FROM sales_reps s
5     JOIN accounts a
6     ON a.sales_rep_id = s.id
7     JOIN orders o
8     ON o.account_id = a.id
9     JOIN region r
10    ON r.id = s.region_id
11    GROUP BY 1,2
12    ORDER BY 3 DESC),
13 t2 AS (
14     SELECT region_name, MAX(total_amt) total_amt
15     FROM t1
16     GROUP BY 1)
17     SELECT t1.rep_name, t1.region_name, t1.total_amt
18     FROM t1
19     JOIN t2
20     ON t1.region_name = t2.region_name AND t1.total_amt =
21           t2.total_amt;
```

Success!

EVALUATE

Output 4 results

Download CSV

rep_name	region_name	total_amt
Earlie Schleusner	Southeast	1098137.72
Tia Amato	Northeast	1010690.60
Georgianna Chisholm	West	886244.12
Charles Bidwell	Midwest	675637.19

2. For the region with the largest sales **total_amt_usd**, how many **total** orders were placed?

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 WITH t1 AS (
2     SELECT r.name region_name, SUM(o.total_amt_usd)
3         total_amt
4     FROM sales_reps s
5     JOIN accounts a
6         ON a.sales_rep_id = s.id
7     JOIN orders o
8         ON o.account_id = a.id
9     JOIN region r
10        ON r.id = s.region_id
11    GROUP BY r.name),
12 t2 AS (
13     SELECT MAX(total_amt)
14     FROM t1)
15     SELECT r.name, SUM(o.total) total_orders
16     FROM sales_reps s
17     JOIN accounts a
18         ON a.sales_rep_id = s.id
19     JOIN orders o
20         ON o.account_id = a.id
21     JOIN region r
22         ON r.id = s.region_id
23     GROUP BY r.name
24 HAVING SUM(o.total_amt_usd) = (SELECT * FROM t2);
```

Success!

EVALUATE

Output 1 results

Download CSV

name	total_orders
Northeast	1230378

3. For the **name** of the account that purchased the most (in total over their lifetime as a customer) **standard_qty**, **how many accounts** still had more in **total** purchases?

Input

```
1 WITH t1 AS (
2     SELECT a.name account_name, SUM(o.standard_qty)
3         total_std, SUM(o.total) total
4     FROM accounts a
5     JOIN orders o
6     ON o.account_id = a.id
7     GROUP BY 1
8     ORDER BY 2 DESC
9     LIMIT 1),
10    t2 AS (
11        SELECT a.name
12        FROM orders o
13        JOIN accounts a
14        ON a.id = o.account_id
15        GROUP BY 1
16        HAVING SUM(o.total) > (SELECT total FROM t1))
17    SELECT COUNT(*)
18    FROM t2;
```

Success!

EVALUATE

Output 1 results

Download CSV

count
3

4. For the customer that spent the most (in total over their lifetime as a customer) **total_amt_usd**, how many **web_events** did they have for each channel?

Input

```

1 WITH t1 AS (
2   SELECT a.id, a.name, SUM(o.total_amt_usd)
3     tot_spent
4   FROM orders o
5   JOIN accounts a
6     ON a.id = o.account_id
7   GROUP BY a.id, a.name
8   ORDER BY 3 DESC
9   LIMIT 1)
10  SELECT a.name, w.channel, COUNT(*)
11  FROM accounts a
12  JOIN web_events w
13  ON a.id = w.account_id AND a.id = (SELECT id FROM
14    t1)
15  GROUP BY 1, 2
16  ORDER BY 3 DESC;

```

Success! **EVALUATE**

Output 6 results [Download CSV](#)

name	channel	count
EOG Resources	direct	44
EOG Resources	organic	13
EOG Resources	adwords	12
EOG Resources	facebook	11
EOG Resources	twitter	5
EOG Resources	banner	4

5. What is the lifetime average amount spent in terms of **total_amt_usd** for the top 10 total spending **accounts**?

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 WITH t1 AS (
2     SELECT a.id, a.name, SUM(o.total_amt_usd)
3         tot_spent
4     FROM orders o
5     JOIN accounts a
6     ON a.id = o.account_id
7     GROUP BY a.id, a.name
8     ORDER BY 3 DESC
9     LIMIT 10)
10 SELECT AVG(tot_spent)
    FROM t1;
```

Success! EVALUATE

Output 1 results Download CSV

avg
304846.969000000000

6. What is the lifetime average amount spent in terms of **total_amt_usd** for only the companies that spent more than the average of all **accounts**.

The screenshot shows a database query editor interface. On the left, there's a schema browser with tables: accounts, orders, region, sales_reps, and web_events. The main area contains a SQL query:

```
1  WITH t1 AS (
2      SELECT AVG(o.total_amt_usd) avg_all
3      FROM orders o
4      JOIN accounts a
5      ON a.id = o.account_id),
6  t2 AS (
7      SELECT o.account_id, AVG(o.total_amt_usd) avg_amt
8      FROM orders o
9      GROUP BY 1
10     HAVING AVG(o.total_amt_usd) > (SELECT * FROM t1))
11     SELECT AVG(avg_amt)
12     FROM t2;
```

The status bar at the bottom says "Success!" and there is an "EVALUATE" button. Below the query results, it says "Output 1 results" and "avg". The result value is 4721.1397439971747168.

QUIZ QUESTION

Look back through your solutions to subqueries and CTEs (WITH statements), which do you find more readable? Which is better for performance?

- CTEs are more readable.
- CTEs are more efficient, as the tables aren't recreated with each subquery portion.

Note: Perhaps you find Subqueries more readable, but most find CTEs to be better for both of these options. They are certainly more efficient from a computing perspective.

Subquery Conclusion

In this lesson, you learned how to write subqueries and use temp tables. Writing queries like the ones you learned here, allows you to substantially increase the types of problems you can solve and the volume of data you can comfortably work with in SQL.

In the next lesson, you'll learn a few tools to handle more directed use cases.

Recap

This lesson was the first of the more advanced sequence in writing SQL. Arguably, the advanced features of **Subqueries** and **CTEs** are the most widely used in an analytics role within a company. Being able to break a problem down into the necessary tables and finding a solution using the resulting table is very useful in practice.

If you didn't get the solutions to these queries on the first pass, don't be afraid to come back another time and give them another try. Additionally, you might try coming up with some questions of your own to see if you can find the solution.

The remaining portions of this course may be key to certain analytics roles, but you have now covered all of the main SQL topics you are likely to use on a day to day basis.

Introduction to SQL Data Cleaning

In this lesson, you will be learning a number of techniques to

1. Clean and re-structure messy data.
2. Convert columns to different data types.
3. Tricks for manipulating **NULLs**.

This will give you a robust toolkit to get from raw data to clean data that's useful for analysis.

It's often the case, The data isn't structured exactly as you'd like in a couple of lessons you worked with the web events table, which contains a refer URL that's well pretty messy . In this lesson, you'll learn a number of techniques for cleaning and reorganizing messy data, so that you can group it the way you want . You'll also learn how to change your columns data type and some tricks for how to manipulate nulls. Put all that together, and you'll have a robust toolset for getting from row data to something that's very useful for analysis.

LEFT & RIGHT

Here we looked at three new functions:

1. LEFT
2. RIGHT
3. LENGTH

LEFT pulls a specified number of characters for each row in a specified column starting at the beginning (or from the left). As you saw here, you can pull the first three digits of a phone number using **LEFT(phone_number, 3)**.

RIGHT pulls a specified number of characters for each row in a specified column starting at the end (or from the right). As you saw here, you can pull the last eight digits of a phone number using **RIGHT(phone_number, 8)**.

LENGTH provides the number of characters for each row of a specified column. Here, you saw that we could use this to get the length of each phone number as **LENGTH(phone_number)**.

We'll start by looking at this data set. It's a list of leads for Porch and Posy, Prospective customers whose business that they might try to win. Please keep in mind that this isn't anybody's real personal information. This was randomly generated for the sake of this lesson.

Our goal here is to clean up this data set, to make it more useful for analysis. For text fields, that means making clean groups that will be useful to aggregate across.

Let's start by pulling the area code out of the phone number. Since the structure of a phone number in this data set is always the same, we can use that to our advantage. The area code is always the first three characters of a phone number. We can get this using a left function. You can use left to pull certain number of characters from the left side of the string, and present them as a separate string.

The screenshot shows a SQL query editor interface with the following details:

- Toolbar:** Run, Limit 100, Format SQL, View History...
- Query:**

```

1 SELECT first_name,
2      last_name,
3      phone_number,
4      LEFT(phone_number, 3) AS area_code,
5      FROM demo.customer_data
    
```
- Status Bar:** ✓ Succeeded in 1s
- Table Headers:** Export, Copy, Chart, Pivot, 1000 rows returned
- Data Table:**

	first_name	last_name	phone_number	area_code
1	Alric	Gouny	399-751-5387	399
2	Thatcher	Buscher	711-549-5882	711
3	Zak	Gabby	124-829-9663	124
4	Sheppard	Gatty	216-394-4468	216
5	Clayborn	Gethyn	769-454-6689	769
6	Zorina	Goggen	408-635-6341	408

If we just want the phone number, we can use right, which does the same thing but from the right side. Right works well in this case, because we know that the number of characters will be consistent across the entire phone number field.

The screenshot shows a SQL query editor interface with the following details:

- Toolbar:** Run, Limit 100, Format SQL, View History...
- Query:**

```

1 SELECT first_name,
2      last_name,
3      phone_number,
4      LEFT(phone_number, 3) AS area_code,
5      RIGHT(phone_number, 6) AS phone_number_only,
6      FROM demo.customer_data
    
```
- Status Bar:** ✓ Succeeded in 982ms
- Table Headers:** Export, Copy, Chart, Pivot, 1000 rows returned
- Data Table:**

	first_name	last_name	phone_number	area_code	phone_number_only
1	Alric	Gouny	399-751-5387	399	751-5387
2	Thatcher	Buscher	711-549-5882	711	549-5882
3	Zak	Gabby	124-829-9663	124	829-9663
4	Sheppard	Gatty	216-394-4468	216	394-4468
5	Clayborn	Gethyn	769-454-6689	769	454-6689
6	Zorina	Goggen	408-635-6341	408	635-6341

If it wasn't consistent, it's still possible to pull a string from the right side, in a way that makes sense. The length function returns the length of the string .So, the length of the phone number will always return 12 in this dataset. Since we know that the first three characters will be the area code, and they'll be followed by a dash. So, total of four characters. We could represent the right function as a function of the length.



```

Run Limit 100 Format SQL View History...
1 | SELECT first_name,
2 |     last_name,
3 |     phone_number,
4 |     LEFT(phone_number, 3) AS area_code,
5 |     RIGHT(phone_number, 8) AS phone_number_only,
6 |     RIGHT(phone_number, LENGTH(phone_number) - 4) AS phone_number_alt
7 | FROM demo.customer_data
  ✓ Succeeded in 1s

```

Export Copy Chart Pivot 1000 rows returned

	first_name	last_name	phone_number	area_code	phone_number_only	phone_number_alt
1	Alric	Gouny	399-751-5387	399	751-5387	751-5387
2	Thatcher	Buscher	711-549-5882	711	549-5882	549-5882
3	Zak	Gabby	124-829-9663	124	829-9663	829-9663
4	Sheppard	Gatty	216-394-4468	216	394-4468	394-4468
5	Clayborn	Gethyn	769-454-6689	769	454-6689	454-6689
6	Zorina	Goggen	408-635-6341	408	635-6341	635-6341

As you can see, the results in the phone number alt field are the same as the result in the phone number only field. This is because , length of the phone number minus four is going to resolve as twelve minus four or eight. Which is the same as the phone number only call.

When using functions within other functions, it's important to remember that the innermost functions will be evaluated first, followed by the functions that encapsulate

Quiz: LEFT & RIGHT

LEFT & RIGHT Quizzes

1. In the **accounts** table, there is a column holding the **website** for each company. The last three digits specify what type of web address they are using.

The screenshot shows a database interface with the following details:

Input:

- SCHEMA dropdown menu showing: accounts, orders, region, sales_reps, web_events.
- SQL code entered:

```
1 SELECT RIGHT(website, 3) AS domain, COUNT(*)  
2   num_companies  
3 FROM accounts  
4 GROUP BY 1  
5 ORDER BY 2 DESC;
```
- Status message: Success!
- EVALUATE button.

Output:

- 3 results
- Download CSV button.
- Data table:

domain	num_companies
com	349
net	1
org	1

2. There is much debate about how much the name (or even the first letter of a company name) matters. Use the **accounts** table to pull the first letter of each company name to see the distribution of company names that begin with each letter (or number).

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```

1  SELECT LEFT(UPPER(name), 1) AS first_letter, COUNT(*)
2    num_companies
3   FROM accounts
4  GROUP BY 1
5 ORDER BY 2 DESC;

```
- Status:** Success!
- Output:** 26 results
- Data:**

first_letter	num_companies
C	37
A	37
P	27

3. Use the **accounts** table and a **CASE** statement to create two groups: one group of company names that start with a number and a second group of those company names that start with a letter. What proportion of company names start with a letter?

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events.
- Query:**

```

1  SELECT SUM(num) nums, SUM(letter) letters
2  FROM (SELECT name, CASE WHEN LEFT(UPPER(name), 1) IN
3        ('0','1','2','3','4','5','6','7','8','9')
4        THEN 1 ELSE 0 END AS num,
5        CASE WHEN LEFT(UPPER(name), 1) IN
6        ('0','1','2','3','4','5','6','7','8','9')
7        THEN 0 ELSE 1 END AS letter
8     FROM accounts) t1;

```
- Status:** Success!
- Output:** 1 results
- Data:**

nums	letters
1	350

There are 350 company names that start with a letter and 1 that starts with a number. This gives a ratio of 350/351 that are company names that start with a letter or 99.7%.

4. Consider vowels as a, e, i, o, and u. What proportion of company names start with a vowel, and what percent start with anything else?

The screenshot shows a database interface with the following details:

- Input Schema:** accounts, orders, region, sales_reps, web_events
- SQL Query:**

```
1 SELECT SUM(vowels) vowels, SUM(other) other
2 FROM (SELECT name, CASE WHEN LEFT(UPPER(name), 1) IN
3 ('A','E','I','O','U')
4 THEN 1 ELSE 0 END AS vowels,
5 CASE WHEN LEFT(UPPER(name), 1) IN
6 ('A','E','I','O','U')
7 THEN 0 ELSE 1 END AS other
8 FROM accounts) t1;
```
- Status:** Success!
- Output:** 1 results
- Data:**

vowels	other
80	271
- Buttons:** EVALUATE, Download CSV

There are 80 company names that start with a vowel and 271 that start with other characters. Therefore 80/351 are vowels or 22.8%. Therefore, 77.2% of company names do not start with vowels.

POSITION, STRPOS, & SUBSTR

In this lesson, you learned about:

1. POSITION
2. STRPOS
3. LOWER
4. UPPER

POSITION takes a character and a column, and provides the index where that character is for each row. The index of the first position is 1 in SQL. If you come from another programming language, many begin indexing at 0. Here, you saw that you can pull the index of a comma as **POSITION(',') IN city_state)**.

STRPOS provides the same result as **POSITION**, but the syntax for achieving those results is a bit different as shown here: **STRPOS(city_state, ',')**.

Note, both **POSITION** and **STRPOS** are case sensitive, so looking for **A** is different than looking for **a**.

Therefore, if you want to pull an index regardless of the case of a letter, you might want to use **LOWER** or **UPPER** to make all of the characters lower or uppercase.

Left and right work pretty well in specific circumstances. When the date is structured very cleanly with a certain number of characters. If you want a separate city and state, you've got to do a little more work. The first thing that will be helpful is figuring out exactly where the city and state split. Since it will be different for each row, we have to use a function that will find the comma and identify how far into the record it is.

Position allows you to specify a sub-string, then it returns a numerical value equal to how far away from the left that particular character appears.

You can also use the string position function which is annotated as STRPOS to achieve the same results. just replace in with a comma and switch the order of the string and the sub-string.

The screenshot shows a SQL query editor interface. At the top, there are buttons for 'Run' (highlighted in green), 'Limit 100', 'Format SQL', and 'View History...'. Below the buttons is the SQL code:

```
1 SELECT first_name,
2      last_name,
3      city_state,
4      POSITION(, ' IN city_state) AS comma_position
5 FROM demo.customer_data
```

In the bottom right corner of the editor area, there is a message: '✓ Succeeded in 954ms'. Below the editor, there is a toolbar with 'Export', 'Copy', 'Chart', and 'Pivot' options. To the right of the toolbar, it says '1000 rows returned' with a refresh icon. The main area displays a table with the following data:

	first_name	last_name	city_state	comma_position
1	Alric	Gouny	Cincinnati, OH	11
2	Thatcher	Buscher	Anchorage, AK	10
3	Zak	Gabby	Saginaw, MI	8
4	Sheppard	Gatty	Jacksonville, FL	13
5	Clayborn	Gethyn	San Antonio, TX	12
6	Zorina	Goggen	Alpharetta, GA	11

Importantly, both the position and string position functions are case sensitive. If you want to look for a character regardless of its case, you can make the entire string upper or lower case.

You can use LOWER to force every character in a string to become lowercase. similarly, you can use UPPER to make all the letters appear as uppercase. Let's complete the loop and pull just the city out into its own field. We can do this by nesting the position inside a LEFT function.

The screenshot shows a SQL query editor interface. At the top, there are buttons for 'Run' (highlighted in green), 'Limit 100', 'Format SQL', and 'View History...'. Below the buttons is the SQL code:

```

1 SELECT first_name,
2       last_name,
3       city_state,
4       POSITION(',') IN city_state) AS comma_position,
5       STRPOS(city_state,',') AS substr_comma_position,
6       LOWER(city_state) AS lowercase,
7       UPPER(city_state) AS uppercase
8   FROM demo.customer_data

```

In the bottom right corner of the editor window, there is a message: '✓ Succeeded in 1s'.

Below the editor, there is a navigation bar with links: 'Export', 'Copy', 'Chart', 'Pivot'. To the right of the navigation bar, it says '1000 rows returned'.

	first_name	last_name	city_state	comma_position	substr_comma_position	lowercase	uppercase
1	Alric	Gouny	Cincinnati, OH	11	11	cincinnati, oh	CINCINNATI, OH
2	Thatcher	Buscher	Anchorage, AK	10	10	anchorage, ak	ANCHORAGE, AK
3	Zak	Gabby	Saginaw, MI	8	8	saginaw, mi	SAGINAW, MI
4	Sheppard	Gatty	Jacksonville, FL	13	13	jacksonville, fl	JACKSONVILLE, FL
5	Clayborn	Gethyn	San Antonio, TX	12	12	san antonio, tx	SAN ANTONIO, TX
6	Zorina	Goggen	Alpharetta, GA	11	11	alpharetta, ga	ALPHARETTA, GA

As you can see, what we really want isn't the full text up to the position of the comma, we want to end one position before the comma so that it's not included in our city column. We can do this by subtracting one within the left function.

Quiz: POSITION, STRPOS, & SUBSTR – AME DATA AS QUIZ 1

You will need to use what you have learned about **LEFT** & **RIGHT**, as well as what you know about **POSITION** or **STRPOS** to do the following quizzes.

1. Use the **accounts** table to create **first** and **last** name columns that hold the first and last names for the **primary_poc**.

The screenshot shows a database interface with the following details:

- Input Schema:** The user has selected the **accounts** schema.
- Query:** The user has entered the following SQL code:

```
1 SELECT LEFT(primary_poc, STRPOS(primary_poc, ' ') - 1 )  
2           first_name,  
3           RIGHT(primary_poc, LENGTH(primary_poc) -  
4           STRPOS(primary_poc, ' ')) last_name  
5     FROM accounts;
```
- Result:** The query was successful, indicated by the green "Success!" message.
- EVALUATE:** A blue button labeled "EVALUATE" is visible.
- Output:** The results show 351 rows with two columns: **first_name** and **last_name**. The data includes:

first_name	last_name
Tamara	Tuma
Sung	Shields
Jodee	Lupo
Serafina	Banda
Angeles	Crusoe
Savanna	Gayman
- Download CSV:** A link to download the results as a CSV file is available.

2. Now see if you can do the same thing for every rep `name` in the `sales_rep` table. Again provide **first** and **last** name columns.

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

HISTORY ▾ MENU ▾

```
1  SELECT LEFT(name, STRPOS(name, ' ') -1 ) first_name,
2      RIGHT(name, LENGTH(name) - STRPOS(name, ' '))
3  FROM sales_reps;
```

Success!

EVALUATE

Output 50 results

Download CSV

first_name	last_name
Samuel	Racine
Eugena	Esser
Michel	Averette
Renetta	Carew
Cara	Clarke
Lavera	Oles

CONCAT

In this lesson you learned about:

1. CONCAT
2. Piping ||

Each of these will allow you to combine columns together across rows. In this video, you saw how first and last names stored in separate columns could be combined together to create a full name: **CONCAT(first_name, '', last_name)** or with piping as **first_name || '' || last_name**.

Now let's say we want to combine first and last names into a single full name column. You can combine strings from several columns using CONCAT. Simply order the values you want to concatenate and separate them with commas. If you want a hard code values, enclose them in single quotes the way that I'm doing with this space.

The screenshot shows a SQL query editor interface. At the top, there are buttons for 'Run' (highlighted in green), 'Limit 100', 'Format SQL', and 'View History...'. Below the buttons is the SQL code:

```
1 SELECT first_name,
2      last_name,
3      CONCAT(first_name, ' ',last_name) AS full_name
4 FROM demo.customer_data
```

On the right side of the editor, a message box indicates: "✓ Succeeded in 946ms". Below the editor is a table preview showing the results of the query:

	first_name	last_name	full_name
1	Alric	Gouny	Alric Gouny
2	Thatcher	Buscher	Thatcher Buscher
3	Zak	Gabby	Zak Gabby
4	Sheppard	Gatty	Sheppard Gatty
5	Clayborn	Gethyn	Clayborn Gethyn
6	Zorina	Goggen	Zorina Goggen

At the bottom of the table preview, it says "1000 rows returned".

Alternatively, you can use two pipe characters to perform the same concatenation.

The screenshot shows a SQL query editor interface. At the top, there are buttons for 'Run' (highlighted in green), 'Limit 100', 'Format SQL', and 'View History...'. Below the buttons is the SQL code:

```
1 SELECT first_name,
2       last_name,
3       CONCAT(first_name, ' ', last_name) AS full_name,
4       first_name || ' ' || last_name AS full_name_alt
5 FROM demo.customer_data
```

In the bottom right corner of the editor area, there is a green success message: '✓ Succeeded in 932ms'. Below the editor, there are navigation buttons: 'Export', 'Copy', 'Chart', and 'Pivot'. To the right of these buttons, it says '1000 rows returned' with a refresh icon. The main area displays a table with the following data:

	first_name	last_name	full_name	full_name_alt
1	Airic	Gouny	Airic Gouny	Airic Gouny
2	Thatcher	Buscher	Thatcher Buscher	Thatcher Buscher
3	Zak	Gabby	Zak Gabby	Zak Gabby
4	Sheppard	Gatty	Sheppard Gatty	Sheppard Gatty
5	Clayborn	Gethyn	Clayborn Gethyn	Clayborn Gethyn
6	Zorina	Goggen	Zorina Goggen	Zorina Goggen
7	Merrill	Etherington	Merrill Etherington	Merrill Etherington
8	Mord	Garlick	Mord Garlick	Mord Garlick
9	Rodrigo	Christoffels	Rodrigo Christoffels	Rodrigo Christoffels

Quiz: CONCAT

Quizzes CONCAT

1. Each company in the `accounts` table wants to create an email address for each `primary_poc`. The email address should be the first name of the `primary_poc` . last name `primary_poc` @ company name `.com`.

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1 WITH t1 AS (
2   SELECT LEFT(primary_poc,      STRPOS(primary_poc, ' ')
3           -1) first_name, RIGHT(primary_poc,
4             LENGTH(primary_poc) - STRPOS(primary_poc, ' '))
5           last_name, name
6   FROM accounts)
7   SELECT first_name, last_name, CONCAT(first_name, '.', 
8           last_name, '@', name, '.com')
9   FROM t1;
```

HISTORY ▾ MENU ▾

Success! EVALUATE

Output 351 results

first_name	last_name	concat
Tamara	Tuma	Tamara.Tuma@Walmart.com
Sung	Shields	Sung.Shields@Exxon Mobil.com
Jodee	Lupo	Jodee.Lupo@Apple.com
Serafina	Banda	Serafina.Banda@Berkshire Hathaway.com
Angeles	Crusoe	Angeles.Crusoe@McKesson.com
Savanna	Gayman	Savanna.Gayman@UnitedHealth Group.com

2. You may have noticed that in the previous solution some of the company names include spaces, which will certainly not work in an email address. See if you can create an email address that will work by removing all of the spaces in the account name, but otherwise your solution should be just as in question 1.

Input

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1 WITH t1 AS (
2   SELECT LEFT(primary_poc,      STRPOS(primary_poc, ' '
3     ') -1 ) first_name,  RIGHT(primary_poc,
4       LENGTH(primary_poc) - STRPOS(primary_poc, ' '))
5         last_name, name
6   FROM accounts)
7   SELECT first_name, last_name, CONCAT(first_name, '.', 
8     last_name, '@', REPLACE(name, ' ', ''), '.com')
9   FROM t1
```

Success!

EVALUATE

Output 351 results

first_name	last_name	concat
Tamara	Tuma	Tamara.Tuma@Walmart.com
Sung	Shields	Sung.Shields@ExxonMobil.com
Jodee	Lupo	Jodee.Lupo@Apple.com
Serafina	Banda	Serafina.Banda@BerkshireHathaway.com
Angeles	Crusoe	Angeles.Crusoe@McKesson.com

3. We would also like to create an initial password, which they will change after their first log in. The first password will be the first letter of the **primary_poc**'s first name (lowercase), then the last letter of their first name (lowercase), the first letter of their last name (lowercase), the last letter of their last name (lowercase), the number of letters in their first name, the number of letters in their last name, and then the name of the company they are working with, all capitalized with no spaces.

Input

```

SCHEMA      5
accounts     ▾
orders       ▾
region       ▾
SCHEMA      5
accounts     ▾
orders       ▾
region       ▾
sales_reps   ▾
web_events   ▾

```

```

1 WITH t1 AS (
2   SELECT LEFT(primary_poc,      STRPOS(primary_poc, ' '
3 ) -1 ) first_name,  RIGHT(primary_poc,
4   LENGTH(primary_poc) - STRPOS(primary_poc, ' '))
5   last_name, name
6   FROM accounts)
7   SELECT first_name, last_name, CONCAT(first_name, '.',
8   last_name, '@', name, '.com'),
9   LEFT(LOWER(first_name), 1) ||
10  RIGHT(LOWER(first_name), 1) || LEFT(LOWER(last_name),
11  1) || RIGHT(LOWER(last_name), 1) ||
12  LENGTH(first_name) || LENGTH(last_name) ||
13  REPLACE(UPPER(name), ' ', '')
14  FROM t1;

```

Success! EVALUATE

Output 351 results

first_name	last_name	concat	?column
Tamara	Tuma	Tamara.Tuma@Walmart.com	tata64W.
Sung	Shields	Sung.Shields@Exxon Mobil.com	sgss47E>
Jodee	Lupo	Jodee.Lupo@Apple.com	jelo54AP
Serafina	Banda	Serafina.Banda@Berkshire Hathaway.com	saba85B
Angeles	Crusoe	Angeles.Crusoe@McKesson.com	asce76M

CAST

In this video, you saw additional functionality for working with dates including:

1. TO_DATE
2. CAST
3. Casting with ::

`DATE_PART('month', TO_DATE(month, 'month'))` here changed a month name into the number associated with that particular month.

Then you can change a string to a date using **CAST**. **CAST** is actually useful to change lots of column types. Commonly you might be doing as you saw here, where you change a **string** to a **date** using **CAST(date_column AS DATE)**. However, you might want to make other changes to your columns in terms of their data types. You can see other examples [here](#).

In this example, you also saw that instead of **CAST(date_column AS DATE)**, you can use **date_column::DATE**.

Expert Tip

Most of the functions presented in this lesson are specific to strings. They won't work with dates, integers or floating-point numbers. However, using any of these functions will automatically change the data to the appropriate type.

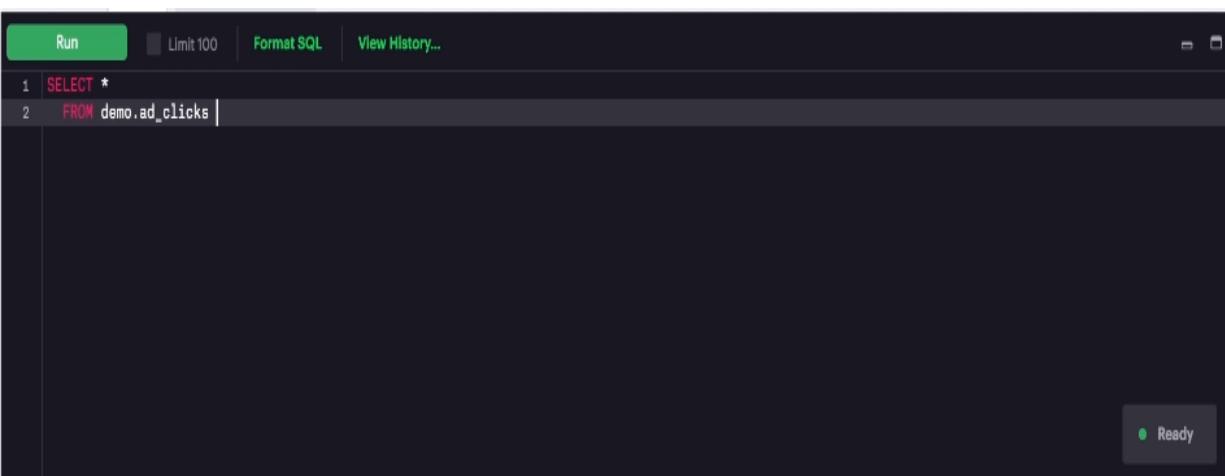
LEFT, **RIGHT**, and **TRIM** are all used to select only certain elements of strings, but using them to select elements of a number or date will treat them as strings for the purpose of the function.

Though we didn't cover **TRIM** in this lesson explicitly, it can be used to remove characters from the beginning and end of a string. This can remove unwanted spaces at the beginning or end of a row that often happen with data being moved from Excel or other storage systems.

There are a number of variations of these functions, as well as several other string functions not covered here. Different databases use subtle variations on these functions, so be sure to look up the appropriate database's syntax if you're connected to a private database. The [Postgres literature](#) contains a lot of the related functions.

We've already explored date functions like date trunk and date part. In an upcoming lesson, we'll dig into even more functions for getting value from dates. The vast majority of analysis you perform in a professional setting will have dates attached to it, but you might not be able to get value out of them, if the dates are not formatted correctly.

Let's say you've got three separate columns for day, month, and year of a given record.



The screenshot shows a SQL query interface with the following details:

- Toolbar buttons: Run, Limit 100, Format SQL, View History...
- SQL Editor:

```
1 SELECT *
2 FROM demo.ad_clicks |
```
- Status Bar: Ready
- Result Table:

	month	day	year	clicks
1	January	1	2014	1135
2	January	2	2014	602
3	January	3	2014	3704
4	January	4	2014	8781
5	January	5	2014	1021
- Bottom Navigation: Export, Copy, Chart, Pivot
- Bottom Status: 1008 rows returned

Problems like this are common for data that was manipulated in excel at some point, or exported from a third party system, or even data than was just manually entered.

In order to really maximize value here, you'll need to convert month names into numbers, then concatenate all of these fields together along with hyphens, and then tell the database to understand that the resulting output is a date, which you can do by using the cast function.

Let's tackle the first couple steps of this first.

The screenshot shows a SQL query editor interface. At the top, there are buttons for 'Run' (highlighted in green), 'Limit 100', 'Format SQL', and 'View History...'. Below the buttons is the SQL code:

```
1 SELECT *,  
2   DATE_PART('month', TO_DATE(month, 'month')) AS clean_month  
3 FROM demo.ad_clicks
```

On the right side of the editor, a message box indicates: **✓ Succeeded in 851ms**. Below the editor is a table with the following data:

	month	day	year	clicks	clean_month
1	January	1	2014	1135	1
2	January	2	2014	602	1
3	January	3	2014	3704	1
4	January	4	2014	8781	1
5	January	5	2014	1021	1

At the bottom of the table, it says **1008 rows returned**.

This first step converted month names into numbers. As you can see, January comes out as one. Now let's concatenate that value together with the year end the day to create something that looks like a real date. For the last step, we'll cast this new field as a date format.

The screenshot shows a SQL query being run in a database environment. The query is:

```

1 SELECT *,  

2     DATE_PART('month',TO_DATE(month, 'month')) AS clean_month,  

3     year || '-' || DATE_PART('month',TO_DATE(month, 'month')) || '-' || day AS concatenated_date,  

4     CAST(year || '-' || DATE_PART('month',TO_DATE(month, 'month')) || '-' || day AS date) AS formatted_date  

5 FROM demo.ad_clicks

```

The results are displayed in a table with the following columns: month, day, year, clicks, clean_month, concatenated_date, and formatted_date. The data shows five rows for January 2014, with the formatted_date column showing the full date (2014-01-01 to 2014-01-05). A success message at the bottom right indicates "Succeeded in 1s".

	month	day	year	clicks	clean_month	concatenated_date	formatted_date
1	January	1	2014	1135	1	2014-1-1	2014-01-01
2	January	2	2014	602	1	2014-1-2	2014-01-02
3	January	3	2014	3704	1	2014-1-3	2014-01-03
4	January	4	2014	8781	1	2014-1-4	2014-01-04
5	January	5	2014	1021	1	2014-1-5	2014-01-05

You can see here that the database now understands this as a date and has attached the appropriate zeros to these records so that it reads properly. The cast function can be a little tricky to read, but luckily there's also a shorthand. Rather than wrapping all of this in a cast function, we can simply add two colons and then the data type we'd like to cast at the end of this section.

The cast function is most useful for turning strings into numbers or dates.

Typically, if you want to turn a number into a string performing any type of string operation on it like left, right, or substring will automatically cast the data while performing the operation.

Quiz: CAST

For this set of quiz questions, you are going to be working with a single table in the environment below. This is a different dataset than Parch & Posey, as all of the data in that particular dataset were already clean.

Tasks to complete:

1. Write a query to look at the top 10 rows to understand the columns and the raw data in the dataset called `sf_crime_data`.

The screenshot shows a SQL query editor interface. The 'Input' section contains a query:

```
1  SELECT *
2  FROM sf_crime_data
3  LIMIT 10;
```

The 'Output' section displays the results of the query, which are 10 rows from the 'sf_crime_data' table:

incident_num	category	descript	day_of_week
1.40099416E8	VEHICLE THEFT	STOLEN AND RECOVERED VEHICLE	Friday
1.40092426E8	ASSAULT	BATTERY	Friday
1.4009241E8	SUSPICIOUS OCC	SUSPICIOUS OCCURRENCE	Friday
1.40092341E8	OTHER OFFENSES	DRIVERS LICENSE, SUSPENDED OR REVOKED	Friday
1.40092573E8	DRUG/NARCOTIC	POSSESSION OF NARCOTICS PARAPHERNALIA	Friday

2. Remembering back to the lesson on dates, use the **Quiz Question** at the bottom of this page to make sure you remember the format that dates should use in SQL.

yyyy-mm-dd

- 3.** Look at the **date** column in the **sf_crime_data** table. Notice the date is not in the correct format.

The format of the **date** column is **mm/dd/yyyy** with times that are not correct also at the end of the date.

- 4.** Write a query to change the date into the correct SQL date format. You will need to use at least **SUBSTR** and **CONCAT** to perform this operation.

The screenshot shows a database query interface. In the 'Input' section, the schema is set to 'sf_crime_data'. The query entered is:

```
1 SELECT date orig_date, (SUBSTR(date, 7, 4) || '-' ||  
2 LEFT(date, 2) || '-' || SUBSTR(date, 4, 2)) new_date  
2 FROM sf_crime_data;
```

The status bar indicates 'Success!' and there is a blue 'EVALUATE' button. In the 'Output' section, it shows 30400 results with two columns: 'orig_date' and 'new_date'. The data is as follows:

orig_date	new_date
01/31/2014 08:00:00 AM +0000	2014-01-31

There is also a 'Download CSV' link.

5. Once you have created a column in the correct format, use either **CAST** or **::** to convert this to a date.

Notice, this new date can be operated on using **DATE_TRUNC**.

The screenshot shows a SQL query editor interface. In the top left, it says "Input" and "SCHEMA sf_crime_data". Below that is the SQL code:

```
1  SELECT date orig_date, (SUBSTR(date, 7, 4) || '-' ||  
2    LEFT(date, 2) || '-' || SUBSTR(date, 4, 2))::DATE  
3    new_date  
4  FROM sf_crime_data;
```

In the bottom right corner of the code area, there is a green "Success!" message and a blue "EVALUATE" button. Below the code area, it says "Output 30400 results" and "Download CSV". The output table has two columns: "orig_date" and "new_date". The first few rows of the output are:

orig_date	new_date
01/31/2014 08:00:00 AM +0000	2014-01-31T00:00:00.000Z

QUIZ QUESTION

What is the correct format of dates in SQL?

dd-mm-yyyy

mm-dd-yyyy

yyyy-mm-dd

yyyy-dd-mm

COALESCE

In this lesson, you learned about how to use **COALESCE** to work with NULL values. Unfortunately, our dataset does not have the **NULL** values that were fabricated in this dataset, so you will work through a different example in the next concept to get used to the **COALESCE** function.

In general, **COALESCE** returns the first non-NULL value passed for each row. Hence why the lesson used **no_poc** if the value in the row was NULL.

Occasionally, you'll end up with a data set that has some nulls that you'd prefer to contain actual values. Looking at the accounts table, you might want to clearly label a no primary point of contact as no POC, so the results will be easily understandable. In cases like this, you can use coalesce to replace the null values.

The screenshot shows a SQL query editor interface. At the top, there are buttons for 'Run', 'Limit 100', 'Format SQL', and 'View History...'. The query itself is:

```
1 SELECT *,  
2 COALESCE(primary_poc, 'no POC') AS primary_poc_modified  
3 FROM demo.accounts  
4 WHERE primary_poc IS NULL
```

At the bottom right of the editor, a green success message says '✓ Succeeded in 1s'. Below the editor, the results are displayed in a table:

	id	name	website	lat	long	primary_poc	sales_rep_id	primary_poc_modified
1	1501	Intel	www.intel.com	41.03153857	-74.66846407	321580	no POC	no POC
2	1671	Delta Air Lines	www.delta.com	40.75860903	-73.99067048	321510	no POC	no POC
3	1951	Twenty-First Century Fox	www.21cf.com	42.35467661	-71.05476697	321560	no POC	no POC
4	2131	USAA	www.usaa.com	41.87745439	-87.62793161	321780	no POC	no POC
5	2141	Duke Energy	www.duke-energy.com	41.87750558	-87.62754203	321790	no POC	no POC

This is something you may want to do frequently when using numerical data where you might want to display nulls as zero. Also, when performing outer joins that result in some unmatched rows, you may want those unmatched rows to display something other than a null value. Of course, this is most valuable when working with a function that treats nulls differently from zero such as a count or an average. We can demonstrate this by wrapping the coalesce in a count function and counting the primary POC column without the coalesce, as well.

The screenshot shows a SQL query editor interface. At the top, there are buttons for 'Run' (highlighted in green), 'Limit 100', 'Format SQL', and 'View History...'. Below the buttons is the SQL code:

```
1 SELECT COUNT(primary_poc) AS regular_count,
2       COUNT(COALESCE(primary_poc, 'no POC')) AS modified_count
3   FROM demo.accounts
```

On the right side of the editor, a status message says '✓ Succeeded in 891ms'. Below the editor, there are buttons for 'Export', 'Copy', 'Chart', and 'Pivot'. To the right, it says '1 rows returned' with a small table icon. The results table has two columns: 'regular_count' and 'modified_count'. There is one row with the value '345' under 'regular_count' and '354' under 'modified_count'.

	regular_count	modified_count
1	345	354

As you can see, the count that includes the coalesce, includes nine more results than the other regular

Quiz: COALESCE

In this quiz, we will walk through the previous example using the following task list.

Tasks to complete:

1. Run the query below to notice the row with missing data.

Input

HISTORY ▾ MENU ▾

SCHEMA	5
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1  SELECT *
2  FROM accounts a
3  LEFT JOIN orders o
4  ON a.id = o.account_id
5  WHERE o.total IS NULL;
```

Success! EVALUATE

Output 1 results [Download CSV](#)

id	name	website	lat	long	primary_poc
	Goldman Sachs Group	www.gs.com	40.75744399	-73.96730918	Loris Manfredi

2. Use **COALESCE** to fill in the `accounts.id` column with the `account.id` for the NULL value for the table in 1.

Input

SCHEMA

accounts

orders

region

sales_reps

web_events

```
1 SELECT COALESCE(a.id, o.id) filled_id, a.name,
2     a.website, a.lat, a.long, a.primary_poc,
3     a.sales_rep_id, o.*
4 FROM accounts a
5 LEFT JOIN orders o
6 ON a.id = o.account_id
7 WHERE o.total IS NULL;
```

Success!

EVALUATE

Output 1 results

Download CSV

filled_id	name	website	lat	long	primary_
1731	Goldman Sachs Group	www.gs.com	40.75744399	-73.96730918	Loris Mar

3. Use **COALESCE** to fill in the `orders.account_id` column with the `account.id` for the NULL value for the table in 1.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1  SELECT COALESCE(a.id, a.id) filled_id, a.name,
2          a.website, a.lat, a.long, a.primary_poc,
3          a.sales_rep_id, COALESCE(o.account_id, a.id)
4          account_id, o.occurred_at, o.standard_qty,
5          o.gloss_qty, o.poster_qty, o.total,
6          o.standard_amt_usd, o.gloss_amt_usd,
7          o.poster_amt_usd, o.total_amt_usd
8
9  FROM accounts a
10
11 LEFT JOIN orders o
12
13 ON a.id = o.account_id
14
15 WHERE o.total IS NULL;
```

Success! EVALUATE

Output 1 results [Download CSV](#)

filled_id	name	website	lat	long	primary_
1731	Goldman Sachs Group	www.gs.com	40.75744399	-73.96730918	Loris Mar

4. Use COALESCE to fill in each of the **qty** and **usd** columns with 0 for the table in 1.

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```

1  SELECT COALESCE(a.id, a.id) filled_id, a.name,
      a.website, a.lat, a.long, a.primary_poc,
      a.sales_rep_id, COALESCE(o.account_id, a.id)
      account_id, o.occurred_at, COALESCE(o.standard_qty,
      0) standard_qty, COALESCE(o.gloss_qty,0) gloss_qty,
      COALESCE(o.poster_qty,0) poster_qty,
      COALESCE(o.total,0) total,
      COALESCE(o.standard_amt_usd,0) standard_amt_usd,
      COALESCE(o.gloss_amt_usd,0) gloss_amt_usd,
      COALESCE(o.poster_amt_usd,0) poster_amt_usd,
      COALESCE(o.total_amt_usd,0) total_amt_usd
2  FROM accounts a
3  LEFT JOIN orders o
4  ON a.id = o.account_id
5  WHERE o.total IS NULL;

```

Success! EVALUATE

Output 1 results [Download CSV](#)

filled_id	name	website	lat	long	primary_
1731	Goldman Sachs Group	www.gs.com	40.75744399	-73.96730918	Loris Mar

5.Run the query in 1 with the **WHERE** removed and **COUNT** the number of **ids** .

Input

HISTORY ▾ MENU ▾

SCHEMA	
accounts	▼
orders	▼
region	▼
sales_reps	▼
web_events	▼

```
1 SELECT COUNT(*)
2 FROM accounts a
3 LEFT JOIN orders o
4 ON a.id = o.account_id;
```

Success!

EVALUATE

Output 1 results

Download CSV

count
6913

6.Run the query in 5, but with the COALESCE function used in questions 2 through 4.

Input

HISTORY ▾ MENU ▾

SCHEMA
accounts
orders
region
sales_reps
web_events

```
1 SELECT COALESCE(a.id, a.id) filled_id, a.name,
a.website, a.lat, a.long, a.primary_poc,
a.sales_rep_id, COALESCE(o.account_id, a.id)
account_id, o.occurred_at, COALESCE(o.standard_qty,
0) standard_qty, COALESCE(o.gloss_qty,0) gloss_qty,
COALESCE(o.poster_qty,0) poster_qty,
COALESCE(o.total,0) total,
COALESCE(o.standard_amt_usd,0) standard_amt_usd,
COALESCE(o.gloss_amt_usd,0) gloss_amt_usd,
COALESCE(o.poster_amt_usd,0) poster_amt_usd,
COALESCE(o.total_amt_usd,0) total_amt_usd
2 FROM accounts a
3 LEFT JOIN orders o
4 ON a.id = o.account_id;
```

Success! EVALUATE

Output 6913 results Download CSV

filled_id	name	website	lat
1001	Walmart	www.walmart.com	40.23849
1001	Walmart	www.walmart.com	40.23849
1001	Walmart	www.walmart.com	40.23849
1001	Walmart	www.walmart.com	40.23849
1001	Walmart	www.walmart.com	40.23849

Recap

You now have a number of tools to assist in cleaning messy data in SQL. Manually cleaning data is tedious, but you now can clean data at scale using your new skills.

For a reminder on any of the data cleaning functionality, the concepts in this lesson are labeled according to the functions you learned.

Memorizing all of this functionality isn't necessary, but you do need to be able to follow documentation, and learn from what you have done in solving previous problems to solve new problems.

إعداد الأستاذة / أميرة عبد الرحمن

amiraabdulrahman38@gmail.com