NULL Values

* Overview:
  + There will always be missing, or “null” values in your data; that’s inevitable. So, it’s important to know how to work with the missing data and what SQL offers to help
  + Before we begin it’s important to understand the difference between null values vs “0”
    - Null represents missing data – the absence of data. There literally is no value available for that row and column.
    - We commonly use 0 to represent the idea of nothing, but it’s still a value
    - The difference is that if we have 0 widgets, we not exactly how many we have: 0. If there is a null value for the number of widgets, we have no idea how many we have. That information doesn’t exist
    - In many cases, null values can be seen as a nuisance. However, when dealing with math, you can use them to your advantage. If a field in a table is optional, it is possible to insert a new record or update a record without adding a value to this field. Then, the field will be saved with a null value.
      * You cannot divide by zero. Just as in Excel, you will get an error.
      * Using the function NULLIF, you can take any zeros and change them to NULL.
      * SQL will recognize this and allow the division to take place.
    - In this lesson we will learn:
      * How to identify null values.
      * Strategies to replace null values.
* A *value* is whatever represents one piece of data
  + If a value is unknown and does not exist in the database, we call it null (a word that means “nothing” or the absence of data
* Not all empty cells are necessarily null. You often need to do a little investigative work to figure out if the missing information is intentionally blank or just unknown (null).
  + In this table, it’s more likely that Meredith’s address is unknown, or null, rather than she doesn’t actually have a home
    - 
    - But in the 3rd record, maybe Samuel just doen’st have a middle name
    - If a cell is left intentionally empty, it’s preferable to use a blank string (double quotes “” or a space “”) rather than a null value
* There are no real null values in a spreadsheet like Excel, just empty cells. But, there’s a trick to find them with SQL
  + Null is a keyword that SQL understands
  + You can select rows that contain no data in a given column by using the IS NULL operator:

SELECT \*

FROM top\_donors

WHERE first IS NULL;

* + - Which means, “SQL, what are the full records for which first names are missing or unknown?”
  + WHERE first = NULL will not work — you can't perform arithmetic on null values.
    - You can use the NOT operator with NULL to identify non-null rows, like so:

SELECT \*

FROM top\_donors

WHERE first IS NOT NULL;

* + - * Which means, “SQL, what are the full records that have first names associated with them?”
* 5 key strategies against the confusion of null values
  + Delete them (this can be risky)
  + Ignore them
  + Guess them
  + Find them out (ask Samuel for his middle name)
  + For numeric fields, fill in the field average
* NULLIF
  + In many cases, nulls can be seen as a nuisance. However, when dealing with math, you can use them to your advantage
    - You cannot divide by zero; just as in Excel, you will get an error
    - Using the function NULLIF, you can take any zeros and change them to null
    - SQL will recognize this and allow the division to take place
* IFNULL
  + IFNULL is the opposite of NULLIF. This function will take a value and turn it into a zero or a null
  + This can be useful when adding and subtracting values, as SQL will not allow you to add or subtract null, but will allow you to add and subtract numbers
* COALESCE
  + Similar to IFNULL and NVL, the COALESCE function returns the first non-null expression among the arguments in its parentheses
    - * Syntax: COALESCE (“expression1”, “expression2”, “another\_expression”, …)
* The NVL, IFNULL, and COALESCE functions can also be used to achieve the same result. IN this case we want null values to be zero.
* What should you do with null values?
  + Remove the rows with null values
    - This can be a good idea in some cases (such as when you have very few null values or you have rows with mostly null values)
  + “Impute” (i.e. replace) the missing values
    - If you have many rows with a small number of null values, you don’t want to throw away all your data. However, null values could also skew your aggregation results.
    - This is where imputation comes in.
      * There are several ways to replace null values. You could replace null values with the mean or median of the missing value’s column
  + Whenever you decide to drop or impute null values, be prepared to defend that decision.
    - There’s not really a right or wrong answer about what you do, but they key is to have a logical reason (or multiple reasons) for why you came up with the decision you made
    - For instance, if there are only four rows with null values in your 10,000-row data set, you could say you just dropped all of the rows with null values because there were so few. But if there were 4,000 rows with null values in your 10,000-row data set, you could say you imputed missing values to preserve the large number of rows with null values.