General Assembly

STUDY GUIDE

HANDLING NULL VALUES

##Values

- A value is a unit of data. Each cell of a spreadsheet or table contains a value. There are both text (or string) values and integers.
- When a value is unknown and does not exist in a database, it is known as "null."
- Do not confuse "null" with "zero." "Zero" is a known value, "null" is an unknown value.
- Not all empty cells are null. You'll need to determine if the cells with missing data were intentionally left blank (the value doesn't exist) or the value is unknown (null). If a cell is left intentionally empty, it's preferable to use a blank string (double quotes ("") or a space in between double quotes (" ") rather than a null value.

##Finding null values

- You can select rows that contain no data in any given column using the IS NULL function.
- You can also use the IS NOT NULL function to identify non-null rows.

##What to Do With Null Values

- Once you've found your null values, you can choose to impute them, exclude them, or work around them.
- **Impute**
    - When you are confident about why the data is missing and you know what the replacement value should be, there are two queries you can use to impute the data, or replace it with substitute values. If the null value is numerical, it's common to replace it with the mean or median of the missing value's column. If it's categorical, the answer may be uncovered through research.
    - Once a new column is created, it contains both original data and imputed values, and it's impossible to tell which is which when looking at the final set.
- **NULLIF**
    - When the data is obviously wrong or doesn't make sense computationally, we can use NULLIF to transform a value to NULL so that it will be skipped over in the computation.
    - The syntax is: NULLIF(expression1, expression2)
- **IFNULL**
    - IFNULL takes a null value and turns it into a zero, which is especially helpful when adding and subtracting values.
        - The syntax is: IFNULL (expression, 0)
- **Exclude**
    - By using a SELECT query in conjunction with the WHERE filter clause,we can eliminate the row of data where the specified column contains missing data, enabling us to perform our analysis without the rows that contain null values.
    - There are some risks to excluding null values.
    - The fact that the data is missing can be indicative of some underlying problem, either with the business or with the data collection process.
    - If a data set has a large amount of missing data, you can lose a lot of valuable information by excluding data, which can impact the significance of your analysis.
- **Work Around**
    - If your data has a number of null values, you can use the COALESCE function to return the first non-null expression among the values within its parentheses:
    - In a COALESCE function, you're defining the hierarchy of columns within your data to return the non-null value in order of importance.