

Using SQL to clean data

Learn basic SQL queries

Transforming data

Weekly challenge 3

🕒 Reading: Glossary Terms and definitions  
5 min

🕒 Quiz: Weekly challenge 3  
9 questions

🎉 Congratulations! You passed!

Grade received 100%

Latest submission received 100%

Quiz + 45 min

Review Learning Objectives

Go to next item

To pass 80% or higher

1. After a company merger, a data analyst receives a dataset with billions of rows of data. They need to leverage this data to identify insights for upper management. What tool would be most efficient for the analyst to use?

👉 Submit my assignment

- ☐ CSV
- ☐ Spreadsheet
- ☒ SQL
- ☐ Word processing

📊 To Pass: 80% or higher

📊 Your grade: 100%

📊 View Feedback

We keep your highest score

2. In which of the following situations would a data analyst use spreadsheets instead of SQL? Select all that apply.

☐ Link ☐ Filter ☒ Export an issue

- ☐ When using a language to interact with multiple database programs
- ☒ When visually inspecting data
- ☒ Correct
- ☐ When working with a dataset with more than 1,000,000 rows
- ☒ When working with a small dataset
- ☒ Correct

3. A data analyst is managing a database of customer information for a retail store. What SQL command can the analyst use to add a new customer to the database?

- ☐ CREATE TABLE IF NOT EXISTS
- ☐ DROP TABLE IF EXISTS
- ☐ UPDATE
- ☒ INSERT INTO
- ☒ Correct

You are working with a database table that contains invoice data. The table includes columns for *invoice\_id* and *billing\_city*. You want to remove duplicate entries for *billing\_city* and sort the results by *invoice\_id*.

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

```
1 SELECT DISTINCT billing_city
2 FROM Invoice
3 ORDER BY invoice_id
```

Run

Reset

```
+++++
| billing_city |
|-----|
| Stuttgart |
| Oslo |
| Brussels |
| Edmonton |
| Boston |
| Frankfurt |
| Berlin |
| Paris |
| Rousseau |
| Dublin |
| London |
| Mountains View |
| Richmond |
| Cupertino |
| Rome |
| Madison |
| Halifax |
| Edinburgh |
| Sydney |
| Santiago |
| Singapore |
| São Paulo |
| Yellowknife |
| Lisbon |
| Amsterdam |
|
+++++
(Output limit exceeded, 53 of 53 total rows shown)
```

What billing city appears in row 15 of your query result?

NOTE: The query index starts at 1 not 0.

- ☐ Santiago  
☐ Oslo  
☒ Reno  
☐ London

 Correct

The clause **DISTINCT** `billing_city` will remove duplicate entries from the `billing_city` column. The complete query is `SELECT DISTINCT billing_city FROM invoice ORDER BY invoice_id`. The **DISTINCT** clause removes duplicate entries from your query result. The billing city Reno appears in row 15 of your query result.

You are working with a database table that contains customer data. The table includes columns about customer location such as `city`, `state`, `country`, and `postal_code`. You want to find what state names are greater than 3 characters.

You write the SQL query below. Add a LENGTH function that will return any state names that are greater than 3 characters long.

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

```
1 SELECT *
2 FROM customer
3 WHERE LENGTH(state) > 3;
```

Run

Reset

customer_id	first_name	last_name	company	address	city	state	country	postal_code	phone
46	Hugh	O'Reilly	None	3 Chatham Street	Dublin	Dublin	Ireland	None	+35312345678

What state is in row 1 of your query result? (Hint: you will have to scroll to the right with your mouse or track pad to locate the indicated column.)

NOTE: The query index starts at 1 not 0.

- ☐ Ireland  
☐ Chile  
☐ India  
☒ Dublin

 **Correct**

The function `LENGTH(state) > 3` will return any state names that are greater than 2 characters long. The complete query is `SELECT * FROM customer WHERE LENGTH(state) > 3`. The `LENGTH` function counts the number of characters a string contains. The country Ireland is in row 1 of your query result.

8. Fill in the blank: \_\_\_\_\_ refers to the process of converting data from one type to another.

- ☒ Typocasting  
☐ Querying  
☐ Formatting  
☐ Cleaning  
☒ Correct

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

- ☐ LENGTH  
☐ TRIM  
☒ CAST  
☐ SUBSTR
- ☒ Correct

4. A data analyst is cleaning survey data. The results for an optional question contain many nulls. What function can the analyst use to eliminate the null values from the results?

- ☐ LENGTH  
☐ CAST  
☐ CONCAT  
☒ COALESCE  
☒ Correct

You are working with a database table that contains customer data. The table includes columns about customer location such as *city*, *state*, and *country*. The state names are abbreviated. You want to retrieve the first 2 letters of each state name. You decide to use the SUBSTR function to retrieve the first 2 letters of each state name, and use the AS command to store the result in a new column called *new\_state*.

You write the SQL query below. Add a statement to your SQL query that will retrieve the first 2 letters of each state name and store the result in a new column as `new_state`.

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

```

1 SELECT
2   customer_id,
3   SUBSTR(state, 1, 2) AS new_state
4 FROM
5   customer
6 ORDER BY
7   state DESC

```

Run

Reset

```

+-----+-----+
| customer_id | new_state |
+-----+-----+
| 29 | MT |
| 17 | WA |
| 48 | WV |
| 28 | UT |
| 26 | TX |
| 1 | SP |
| 18 | SP |
| 11 | SP |
| 47 | IN |
| 12 | NJ |
| 29 | NC |
| 29 | ON |
| 38 | ON |
| 19 | NV |
| 21 | NV |
| 13 | NT |
| 55 | NS |
| 11 | NS |
| 32 | NB |
| 23 | MA |
| 24 | IL |
| 22 | FL |
| 46 | CA |
| 13 | OR |
| 56 | CO |
+-----+-----+
(Output limit exceeded, 35 of 59 total rows shown)

```

What customer ID number is in row 9 of your query result?

NOTE: The query index starts at 1 not 0.

- ☒ 47  
☐ 3  
☐ 10  
☐ 55

 **Correct**

The statement `SUBSTR(state, 1, 2) AS new_state` will retrieve the first 2 letters of each state name and store the result in a new column as `new_state`. The complete query is `SELECT customer_id, SUBSTR(state, 1, 2) AS new_state FROM customer ORDER BY state DESC`. The `SUBSTR` function extracts a substring from a string. This function instructs the database to return 2 characters of each state name, starting with the first character. The customer ID number 47 appears in row 9 of your query result.