SQL & Python for Analytics

USC Annenberg Digital Lounge
Week 2: Data Modeling with SQL

→ Get the guide at: digital-lounge-analytics.carrd.co

Today's session

This is going to be interactive, so please be ready to try out the exercises as we go!

We are going to be using a tool called Hex, which is a tool for using SQL and Python for doing data analysis and visualization. Shout-out to Hex for giving us their Professional Plan for free!

Go to digital-lounge-analytics.carrd.co

Get set up

Go to digital-lounge-analytics.carrd.co and grab the link to Hex.

TODAY'S PROJECT

· Visit this link to sign up for Hex, the free data notebook tool we'll be using during this class: **https://app.hex.tech/ link/1Hw6-74CMkrxy3hLjGqcv8a7**

If asked to select a workspace, choose Annenberg Digital Lounge:

USC Annenberg Digital Lounge
53 members | You are an Editor   →

Get the guide: digital-lounge-analytics.carrd.co

Your workbook

In Hex, look for "**Fall 2025 Week 2**", then click the "**...**" next to the name, then click **Duplicate**.

Fall 2025 Week 2: Data Modeling with SQL

Copy this template to start for Week 2!

- ↗ Open in new tab
- 🔗 Copy link
- ⧉ Duplicate
- ⬆ Export as ▶
- ⚠ Clear outputs
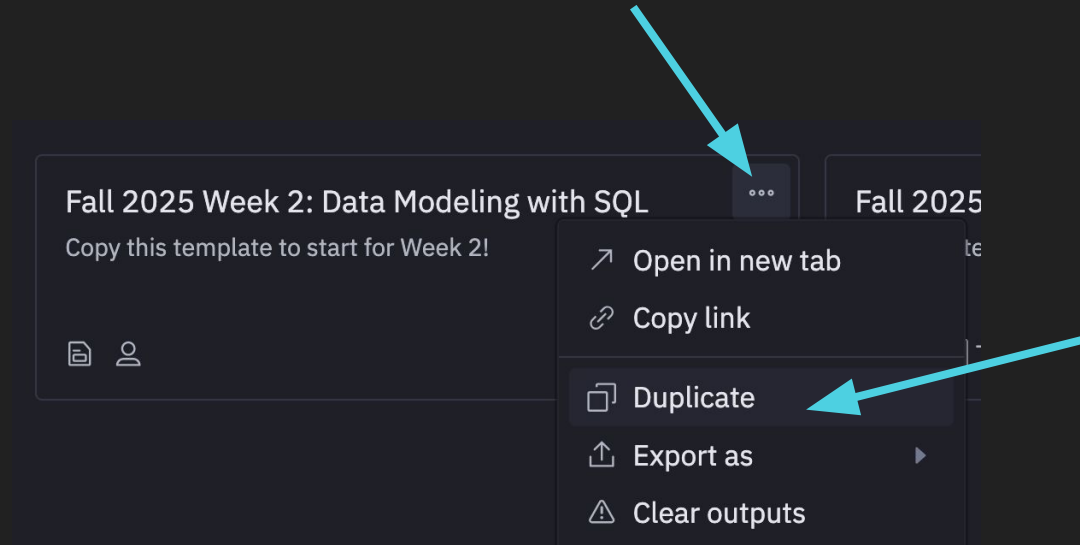
Workshop overview

- Week 1: Intro to SQL
- Week 2 (Today!): Data Modeling with SQL
- Week 3 (Nov. 10): Intro to Python
- Week 4 (Nov. 17): Analyzing Data with Python
- Week 5 (Nov. 24): Analytics & Visualization with SQL and Python

Get the guide: digital-lounge-analytics.carrd.co

Today's agenda

- Review SQL 101
- Data Modeling
- Joining Data Frames
- Casting

SQL Review

SELECT (columns, or a function of columns, or * for all)

FROM (table)

JOIN (other table)

WHERE (conditions)

GROUP BY (columns)

ORDER BY (which column)

LIMIT (if you only want a certain number of results)

Sometimes those words are written in all caps by convention, but it is not required

Get the guide: digital-lounge-analytics.carrd.co

Our First Queries

```sql
select * from transcripts_data limit 20


select distinct character,
count(*) as number_of_lines
from transcripts_data
group by character
order by number_of_lines desc
```

Our First Queries - Formulas

```sql
select season,
"EpisodeNo" as ep_number,
count("Dialogue") as lines_of_dialogue,
count(case when character = 'ELAINE' then "Dialogue" end) as elaine_lines,
count(case when character = 'KRAMER' then "Dialogue" end) as kramer_lines
from transcripts_data
group by season, "EpisodeNo"
```

Our First Queries – Joining Dataframes

```
select * from ratings_data_with_episode_number
join (select * from lines_per_episode) using (season, ep_number)
```

Navigating Hex

- How do I know what columns are available to select from?
- Go to Data Browser → choose a dataframe
- See all the columns listed below

Add SQL cells
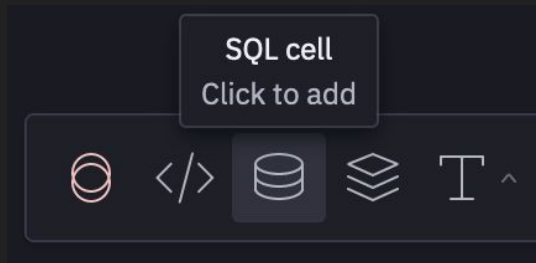
Command-Enter to run a cell (and any dependent cells)

Data browser

Data   Recently used

‹ All sources

Dataframes   ⟋ 5

All dataframes in scope   Query

🔍 Search...

dataframe	136 B
dataframe_2	3.67 KB
dataframe_3	131.37 KB
dataframe_4	136 B
dataframe_5	74.07 KB
ratings_data	74.07 KB
transcripts_data	13.52 MB

ratings_data 6 cols   ✕

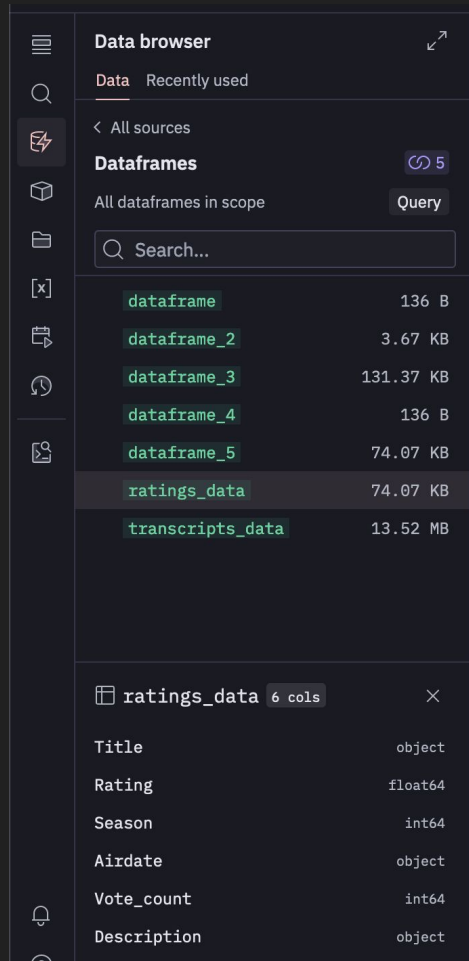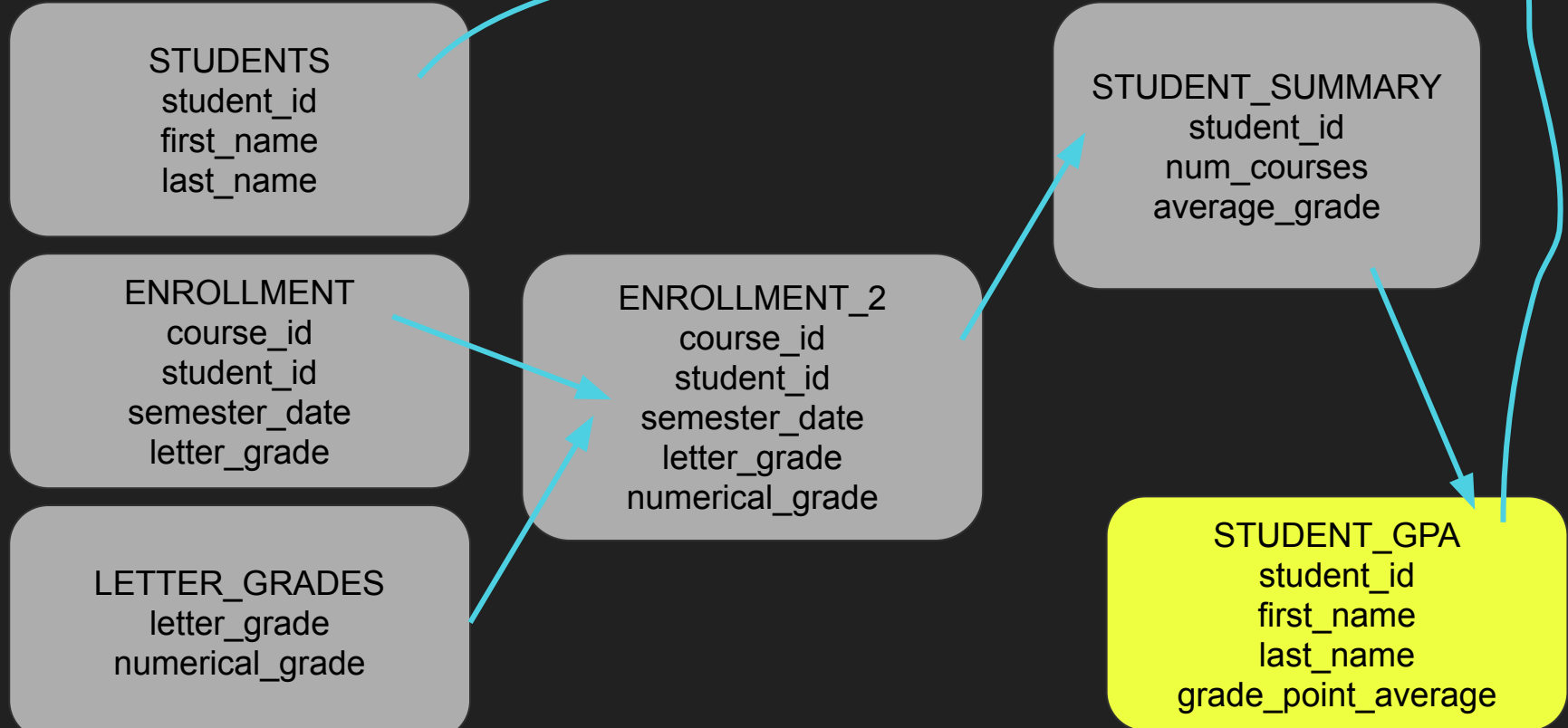| Title | object |
Rating	float64
Season	int64
Airdate	object
Vote_count	int64
Description	object

SQL cell
Click to add

Time for exercise 1!

What is data modeling?

- A set of SQL queries that "clean up" your source data, to make it easy to ask and answer business questions of your data
- Your initial datasets are "source data" – these come from systems
- Your completed data models take that source data, do any joins or formulas that your end users might commonly need to do for their queries

Data model example

STUDENTS
student_id
first_name
last_name

ENROLLMENT
course_id
student_id
semester_date
letter_grade

LETTER_GRADES
letter_grade
numerical_grade

ENROLLMENT_2
course_id
student_id
semester_date
letter_grade
numerical_grade

STUDENT_SUMMARY
student_id
num_courses
average_grade

STUDENT_GPA
student_id
first_name
last_name
grade_point_average

Get the guide: digital-lounge-analytics.carrd.co

Why data modeling?

Get the guide: digital-lounge-analytics.carrd.co

Types of Joins

Regular Join

- All rows from both dataframes that have matching values

- If any row in either dataframe doesn't have a match in the other, it will be dropped

Left Join

- ALL the rows from the left dataframe, plus anything in the right dataframe that has matching values

Steps for joining datasets

Figure out…	Build your query
What tables have the source columns you need?	SELECT those_columns FROM that_table
Does each table have one row per entity? If not, summarize first so that they do in a new dataframe	SELECT sport, count(athlete_id) GROUP BY 1
Do you need only rows that appear in both tables? Or all rows from one table?	JOIN table table_name LEFT JOIN table table_name
For all the tables you need, what column that has a unique identifier that you can use to match records across tables?	USING (same_column_title) ON column1=column2
What columns do you want to end up with? Do you need any formulas?	SELECT those_columns, a_formula(a_column) AS result FROM that_table

Get the guide: digital-lounge-analytics.carrd.co

WHERE Conditions

Can also be used in the WHEN part of CASE statements!

Numbers

- = equal

- != is not equal

- < less than

- > greater than

- <= less than or equal to

- >= greater than or equal to

Strings/Text

- = equal

- != is not equal

- IN ('othertext1', 'othertext2')

- LIKE 'this exact TExt' (case sensitive)

- ILIKE '%this phrase%' (case insensitive)

Time for exercise 2!

Steps for joining datasets

Figure out…	Build your query
What tables have the source columns you need?	SELECT _____ FROM _____
Does each table have one row per entity? If not, summarize first so that they do in a new dataframe	SELECT _____, count(_____) GROUP BY 1
Do you need only rows that appear in both tables? Or all rows from one table?	JOIN _____ LEFT JOIN _____
For all the tables you need, what column that has a unique identifier that you can use to match records across tables?	USING (_____) ON _____=_____
What columns do you want to end up with? Do you need any formulas?	SELECT _____, a_formula(_____) AS result FROM _____

Get the guide: digital-lounge-analytics.carrd.co

Data modeling practice
Exercises 3-9

Casting Data Types

SQL 28 ⚠ SQL 29 ▷ Run all

```
1  -- casting between data types
2
3  select '123' as sample_data
```

1 row 0s 3m ago

≡ Filters   Table

	A sample_data	+
0	123	

↳ **dataframe_14** 1   👁 🖌 ▥ Σ 1 r

SQL 29

```
1  select sum(sample_data) from dataframe_14
```

⚠ **Your query could not be executed**

We received the following error when executing this query:

[◎ Fix with Magic] [Hide error details] [Copy error]

Exception: Binder Error: No function matches the given name and argument types 'sum(VARCHAR)'.
You might need to add explicit type casts.

Casting Data Types


```sql
-- casting between data types

select '123'::int as sample_data
```

Filters

	# sample_data	+
0	123	

↳ **dataframe_14** 1

```sql
select sum(sample_data) from dataframe_14
```

Filters

	# sum(sample_data)	+
0	123	

↳ **dataframe_15**

Get the guide

Keep learning!

Upcoming sessions:

- Week 3 (Nov. 10): Intro to Python
- Week 4 (Nov. 17): Analyzing Data with Python
- Week 5 (Nov. 24): Analytics & Visualization with SQL and Python

Good tutorials at:

- Data modeling: docs.getdbt.com/best-practices/how-we-style/1-how-we-style-our-dbt-models
- learn.hex.tech
- Learn SQL via a Murder Mystery: https://mystery.knightlab.com/
- sqlbolt.com for a more standard tutorial

Get the guide: digital-lounge-analytics.carrd.co