“Height” measured in centimeters is an example of what type of variable?

Correct!

● Continuous

○ Data/time

○ Discrete

○ Nominal

○ Ordinal

Question 2

3 / 3 pts

Is the following dataset “tidy” (Codd’s 3rd normal form), and if it isn’t, what is preventing it from being tidy?

patient	visit	assay	value
1	1	glucose	84
1	1	insulin	0
1	2	glucose	89
1	2	insulin	100
2	1	glucose	134
2	1	insulin	428
2	2	glucose	138
2	2	insulin	202

○ Column names are values rather than variables

Correct!

● Cells encode variable names rather than values

○ Cells encode values for multiple variables

○ The table is "tidy"

○ Single observations are stored in multiple tables

Question 3

3 / 3 pts

“Temperature” measured in {“cold”, “cool”, “moderate”, “warm”, “hot”} is an example of what type of variable?

○ Date/time

○ Continuous

○ Discrete

○ Nominal

Correct! ● Ordinal

Question 4

3 / 3 pts

“Weather” measured in {“sunny”, “cloudy”, “windy”, “rain”, “snow”, “smoke”, “fog”} is an example of what type of variable?

○ Date/time

○ Discrete

Correct! ● Nominal

○ Ordinal

○ Continuous

Question 5

3 / 3 pts

Given the following schema, find all departments that sell at least one item costing greater than $100.

Items(**id**: *string*, description: *string*, dept: *string*, count: *int*, price: *float*)

○ SELECT dept, MAX(price) AS maxprice FROM Items WHERE maxprice > 100 GROUP BY dept

○ SELECT dept, MAX(price) FROM Items GROUP BY dept

○ SELECT dept, SUM(price) FROM Items GROUP BY dept HAVING SUM(price) > 100

○ SELECT dept, MAX(price) AS maxprice FROM Items GROUP BY dept, maxprice > 100

Correct!

● SELECT dept, MAX(price) AS maxprice FROM Items GROUP BY dept HAVING maxprice > 100

Question 6

3 / 3 pts

Which of the following graphics should be avoided due to difficulties in interpretation?

○ Bar plot

○ Scatterplot

Correct!

● Pie chart

○ Boxplot

○ Histogram

Question 7

3 / 3 pts

Which is an example of missing completely at random (MCAR)?

○ Survey responses indicating “Prefer not to say” for many questions

○ Fewer examples of non-white faces in an computer vision training set

○ Lower-income individuals leave “income” question blank

○ Participants drop out of study and do not complete follow-up surveys

Correct!

● Blood samples damaged during transport between laboratories

Question 8

0 / 3 pts

Which is an example of unit non-response?

○ Survey responses indicating “Prefer not to say” for many questions

○ Blood samples damaged during transport between laboratories

orrect Answer

○ Fewer examples of non-white faces in an computer vision training set

○ Lower-income individuals leave “income” question blank ou Answered

● Participants drop out of study and do not complete follow-up surveys

Question 9

3 / 3 pts

Is the following dataset “tidy” (Codd’s 3rd normal form), and if it isn’t, what is preventing it from being tidy?

patient	glucose_visit1	glucose_visit2
1	84	89
2	134	128
3	111	102
4	98	87
5	78	67

Correct!

● Column names are values rather than variables

○ Single observations are stored in multiple tables

○ Cells encode variable names rather than values

○ The table is "tidy"

○ Cells encode values for multiple variables

Question 10

0 / 3 pts

Which is an example of missing at random (MAR)?

○ Fewer examples of non-white faces in an computer vision training set

ou Answered

● Survey responses indicating “Prefer not to say” for many questions orrect Answer ○ Participants drop out of study and do not complete follow-up surveys

○ Blood samples damaged during transport between laboratories

○ Lower-income individuals leave “income” question blank

Question 11

3 / 3 pts

Given the following schema, calculate the average height for trees of each species.

Tree(**id**: *string*, species: *string*, height: *float*, girth: *float*, age: *int*)

○ SELECT species FROM Tree ORDER BY AVG(height)

Correct! ● SELECT species, AVG(height) FROM Tree GROUP BY species

○ SELECT species, SUM(height) FROM Tree ORDER BY species

○ SELECT species, AVG(height) FROM Tree GROUP BY height

○ SELECT species FROM Tree GROUP BY AVG(height)

Question 12

3 / 3 pts

Is the following dataset “tidy” (Codd’s 3rd normal form), and if it isn’t, what is preventing it from being tidy?

tree	month	measure_type	value
A	1	girth	8.3
A	1	height	70.0
A	2	girth	8.6
A	2	height	71.0
B	1	girth	10.5

tree	month	measure_type	value
B	1	height	81.0
B	2	girth	10.7
B	2	height	83.0

○ Cells encode values for multiple variables

○ The table is “tidy”

○ Single observations are stored in multiple tables

Correct!

● Cells encode variable names rather than values

○ Column names are values rather than variables

Question 13

3 / 3 pts

Which of the following graphics is best suited for investigating a relationship between a continous and categorical variable?

○ Bar plot

○ Pie chart

Correct!

● Boxplot

○ Scatterplot

○ Histogram

Question 14

3 / 3 pts

Given the following schema, rank tree species from widest to thinnest based on average girth at age 10.

Tree(**id**: *string*, species: *string*, height: *float*, girth: *float*, age: *int*)

Correct!

● SELECT species, AVG(girth) FROM Tree WHERE age = 10 GROUP BY species ORDER BY AVG(girth) DESC

○ SELECT species, AVG(girth) FROM Tree WHERE age = 10 GROUP BY species ORDER BY AVG(girth)

○ SELECT species, AVG(girth) FROM Tree WHERE age = 10 ORDER BY species, AVG(girth) DESC

○ SELECT species, age = 10, AVG(girth) FROM Tree GROUP BY species ORDER BY girth DESC

○ SELECT species, AVG(girth) FROM Tree WHERE age = 10 GROUP BY species ORDER BY girth DESC

Question 15

3 / 3 pts

Is the following dataset “tidy” (Codd’s 3rd normal form), and if it isn’t, what is preventing it from being tidy?

patient_visit	**glucose**	**insulin**
P1_V1	84	0
P1_V2	89	100
P2_V1	134	428
P2_V2	128	202

patient_visit	glucose	insulin
P3_V1	111	98
P3_V2	102	0

○ The table is "tidy"

Correct! ● Cells encode values for multiple variables

○ Column names are values rather than variables

○ Single observations are stored in multiple tables

○ Cells encode variable names rather than values

Question 16

5 / 5 pts

What are necessary characteristics of a primary key? (choose all that apply)

□ Set of attribute(s) that uniquely identify tuples in another table

Correct! ☑ Cannot include missing values (NULLs)

□ Set of attribute(s) with real-world meaning

□ A foreign key must exist that references it

Correct! ☑ Set of attribute(s) that uniquely identify tuples in this table

Question 17

0 / 5 pts

What are necessary characteristics of a foreign key? (choose all that apply)

Correct! ☑ References a candidate key in another table

☐ It must be a compound key

ou Answered ☑ Set of attribute(s) that uniquely identify tuples in this table

☐ It must be a composite key

orrect Answer ☐ Set of attribute(s) that uniquely identify tuples in another table

Question 18

5 / 5 pts

Which joins preserve all rows from at least one of the tables? (choose all that apply)

☐ SELECT * FROM x, y WHERE x.key=y.key

Correct! ☑ SELECT * FROM x FULL JOIN y ON x.key=y.key

Correct! ☑ SELECT * FROM x LEFT JOIN y ON x.key=y.key

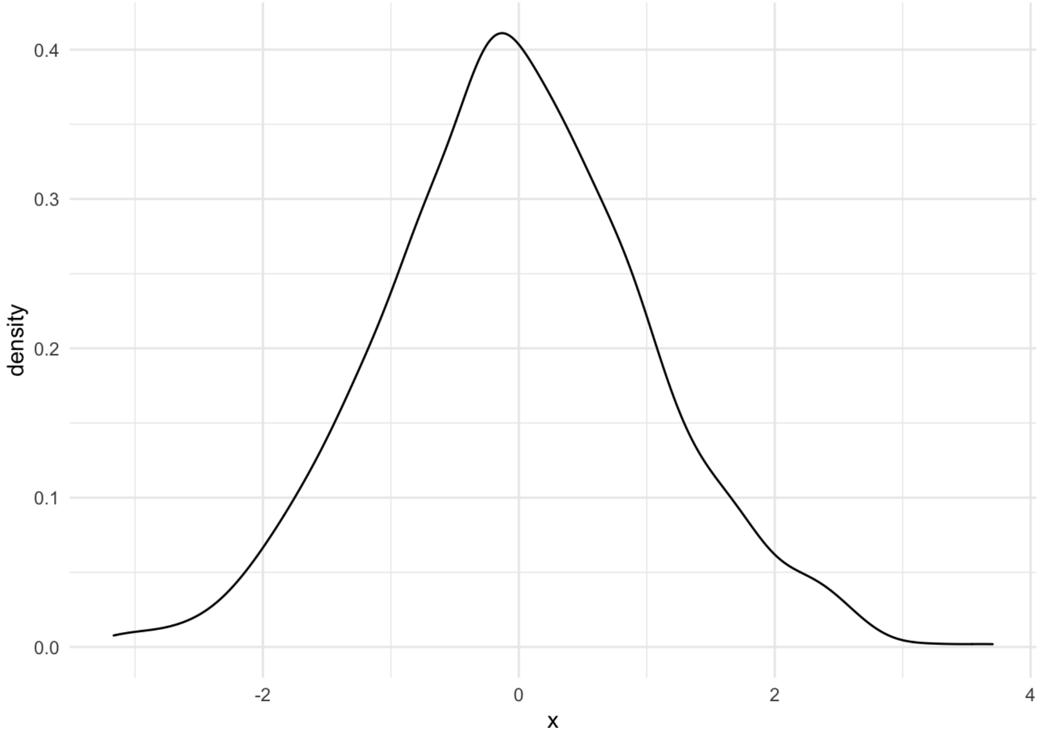
☐ SELECT * FROM x JOIN y ON x.key=y.key

Correct! ☑ SELECT * FROM x RIGHT JOIN y ON x.key=y.key

Question 19

0 / 5 pts

Describe the distribution of the following data (choose all that apply).



□ Uniform

ou Answered

☑ Right skewed

orrect Answer

□ Normal

□ Left skewed

orrect Answer

□ Symmetric

Question 20

5 / 5 pts

Which joins produce output with columns from only the first table? (choose all that apply)

□ Right join

Correct! ☑ Anti join

□ Inner join

□ Left join

Correct! ☑ Semi join

Question 21

5 / 5 pts

What are the reasons for structuring data into a “tidy” format? (choose all that apply)

□ It is efficient for scientific computing (matrix algebra, optimization, etc.)

Correct! ☑ It is easy to query, transform, and aggregate

□ It is the most compact form

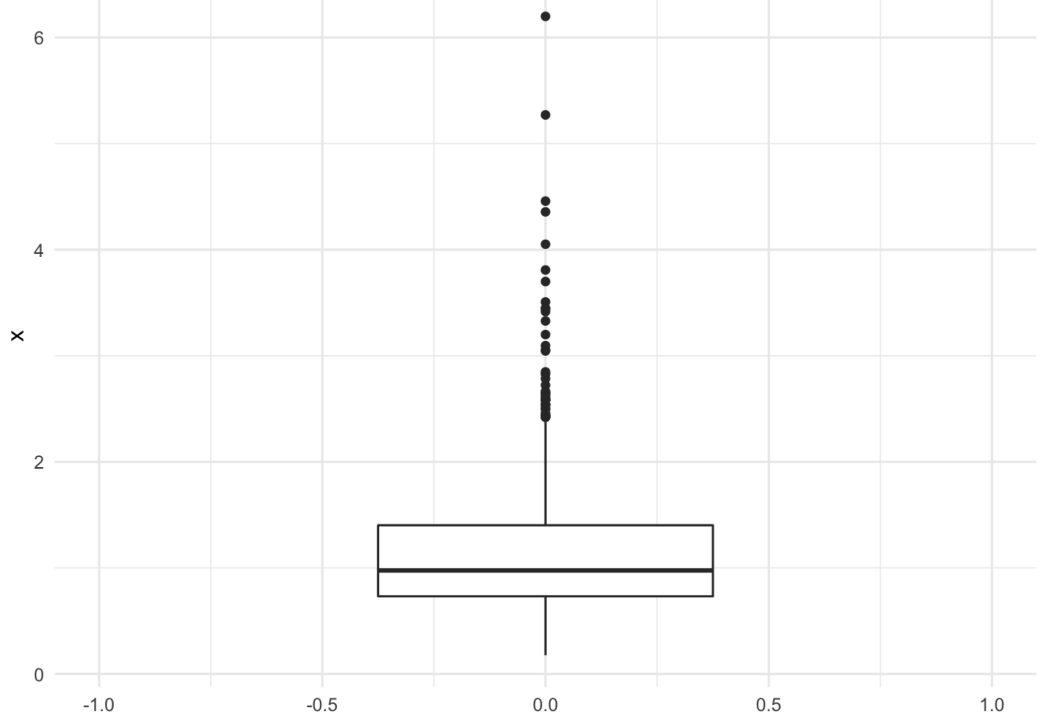
Correct! ☑ It mirrors relational database principles

Correct! ☑ It is easy to visualize using the grammar of graphics

Question 22

5 / 5 pts

Describe the distribution of the following data (choose all that apply).



Correct!

☑ Right skewed

☐ Normal

☐ Uniform

☐ Symmetric

☐ Left skewed

Question 23

0 / 5 pts

What are necessary characteristics of a compound key? (choose all that apply)

☐ It must be a natural key

Correct!

☑ It uses multiple attributes

ou Answered

☑ It consists of multiple candidate keys orrect Answer

□ It consists of foreign keys

□ It must be a surrogate key

Question 24

5 / 5 pts

What are necessary characteristics of “tidy” data? (choose all that apply)

Correct!

☑ Each variable forms a column

Correct!

☑ Each value is a cell

□ Each variable forms a row

□ Each observation forms a column

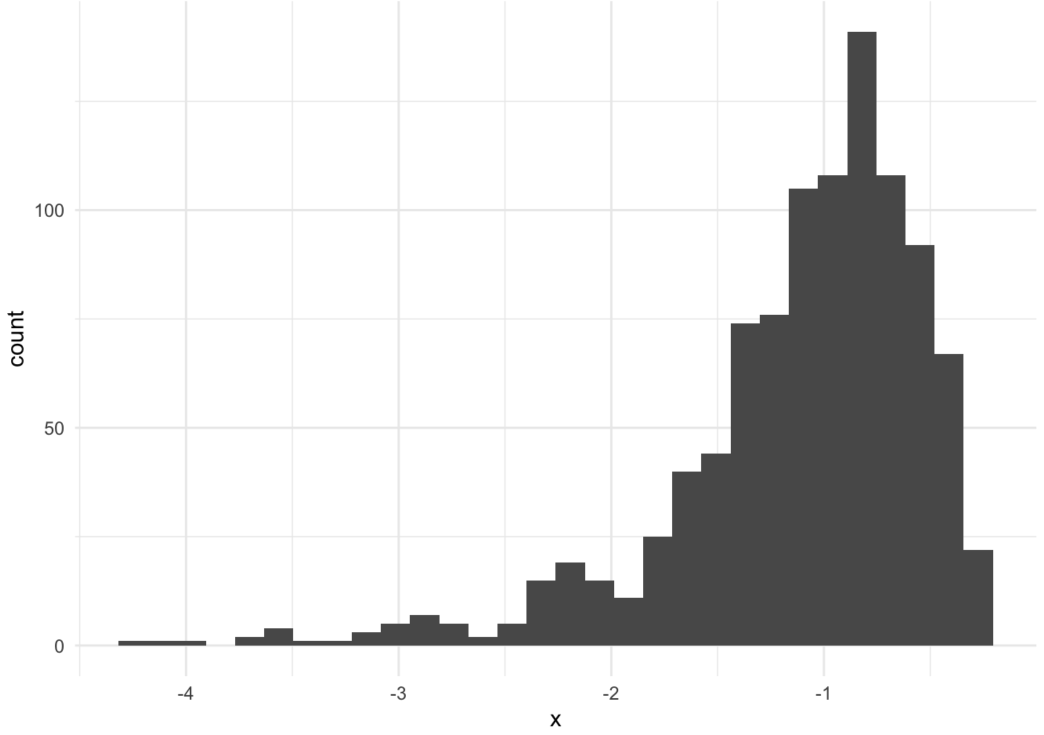
Correct!

☑ Each observation forms a row

Question 25

5 / 5 pts

Describe the distribution of the following data (choose all that apply).



□ Right skewed

□ Uniform

□ Normal

□ Symmetric

Correct!

☑ Left skewed

Question 26

10 / 10 pts

List all foreign keys in the database and the keys they reference, and then identify any compound keys in the database.

The following four data tables describe the available items, orders, users, and deliveries for a certain restaurant’s online delivery service:

items

code	description	price
HAMBGR	Hamburger	4.29
CHZBGR	Cheeseburger	4.99
FRYSML	Small fries	1.99
FRYLRG	Large fries	2.99
FNCOLA	Coke	1.49

orders

order_id	item	quantity
0001	CHZBGR	2
0001	FRYLRG	1
0002	HAMBGR	1
0002	FNCOLA	1
0003	CHZBGR	3

users

login	name	address
john123@geemail.com (mailto:john123@geemail.com)	John	123 Number St
alice89@h0tmail.edu (mailto:alice89@h0tmail.edu)	Alice	88 Infinity Rd
jdepp@pirates.com (mailto:jdepp@pirates.com)	Johnny	21 Jump St
kat22@yah000.net (mailto:kat22@yah000.net)	Kathryn	65 Leonin Ln

deliveries

order_id	user	delivered
0001	alice89@h0tmail.edu (mailto:alice89@h0tmail.edu)	Yes
0002	john123@geemail.com (mailto:john123@geemail.com)	Yes
0003	alice89@h0tmail.edu (mailto:alice89@h0tmail.edu)	No

Your Answer:

foreign keys -

1. orders$item is a foreign key that references items$code

2. orders$order_id is a foreign key that references deliveries$order_id

3. deliveries$user is a foreign key that references users$login

compound keys -

1. orders$order_id, orders$item is a compound key

`orders.order_id` is a foreign key referencing `deliveries.order_id`

`orders.item` is a foreign key referencing `items.code`

`deliveries.user` is a foreign key referencing `users.login`

`(order_id, item)` is a compound primary key for `orders`, because it consists of two foreign keys.

Question 27

9 / 10 pts

Choose a primary key for the `users` table, fully explaining and justifying your answer, including why any other potential candidate keys are not suitable.

The following four data tables describe the available items, orders, users, and deliveries for a certain restaurant's online delivery service:

items

code	description	price
HAMBGR	Hamburger	4.29
CHZBGR	Cheeseburger	4.99
FRYSML	Small fries	1.99
FRYLRG	Large fries	2.99
FNCOLA	Coke	1.49

orders

order_id	item	quantity
0001	CHZBGR	2

order_id	item	quantity
0001	FRYLRG	1
0002	HAMBGR	1
0002	FNCOLA	1
0003	CHZBGR	3

users

login	name	address
john123@geemail.com (mailto:john123@geemail.com)	John	123 Number St
alice89@h0tmail.edu (mailto:alice89@h0tmail.edu)	Alice	88 Infinity Rd
jdepp@pirates.com (mailto:jdepp@pirates.com)	Johnny	21 Jump St
kat22@yah000.net (mailto:kat22@yah000.net)	Kathryn	65 Leonin Ln

deliveries

order_id	user	delivered
0001	alice89@h0tmail.edu (mailto:alice89@h0tmail.edu)	Yes
0002	john123@geemail.com (mailto:john123@geemail.com)	Yes
0003	alice89@h0tmail.edu (mailto:alice89@h0tmail.edu)	No

Your Answer:

login would be the primary key for the users table as each user will have a unique email address.

The other columns i.e. **name** and **address** are not suitable as they might not always be unique. These columns are not convenient to be used as a foreign key in other tables

The `login` attribute is the most appropriate primary key for the given relational database, as it uniquely identifies each row of `users`, and is already used as a foreign key by the `deliveries` table. It is reasonable to guarantee that each user must have a unique email address.

Both `name` and `address` are currently unique, but neither are likely to remain unique in future updates to the table. The combination `(name, address)` is a reasonable candidate key, but it is more complex than `login`, and is not referenced by any other tables.

Should also include that is it used as foreign key in other tables

Question 28

10 / 10 pts

Choose a primary key for the `items` table, fully explaining and justifying your answer, including why any other potential candidate keys are not suitable.

The following four data tables describe the available items, orders, users, and deliveries for a certain restaurant's online delivery service:

items

code	description	price
HAMBGR	Hamburger	4.29
CHZBGR	Cheeseburger	4.99
FRYSML	Small fries	1.99
FRYLRG	Large fries	2.99
FNCOLA	Coke	1.49

orders

order_id	item	quantity
0001	CHZBGR	2
0001	FRYLRG	1
0002	HAMBGR	1
0002	FNCOLA	1
0003	CHZBGR	3

users

login	name	address
john123@geemail.com (mailto:john123@geemail.com)	John	123 Number St
alice89@h0tmail.edu (mailto:alice89@h0tmail.edu)	Alice	88 Infinity Rd
jdepp@pirates.com (mailto:jdepp@pirates.com)	Johnny	21 Jump St
kat22@yah000.net (mailto:kat22@yah000.net)	Kathryn	65 Leonin Ln

deliveries

order_id	user	delivered
0001	alice89@h0tmail.edu (mailto:alice89@h0tmail.edu)	Yes
0002	john123@geemail.com (mailto:john123@geemail.com)	Yes
0003	alice89@h0tmail.edu (mailto:alice89@h0tmail.edu)	No

Your Answer:

code would be the primary key for the ***items*** table as it uniquely identifies each row. And it is easier to use it as a foreign key in other tables

description and **price** of an item can change in the future and cannot be reliable to be considered as a primary key.

The item `code` attribute is the most appropriate primary key for the given relational database, as it uniquely identifies each row of `items`, and is already used as a foreign key by the `orders` table. It can be safely assumed that items will have unique codes.

Both `description` and `price` are currently unique, but neither are likely to remain unique in future updates to the table. Even if `description` remains unique, `code` is a more appropriate primary key as it is already referenced by a foreign key.

Question 29

7 / 10 pts

Provide a SQL-style pseudocode strategy (using relational data concepts such as SELECT, WHERE, GROUP BY, and JOIN) for solving the problem.

Find the number of books from each genre that each agent represents. The resulting table should by sorted by agent, and include agent, genre, and count.

The following three data tables describe the authors/clients, book titles, and sales to publishers for a certain literary agency:

clients

cid	first_name	last_name	sign_date	agent
jsmith	Jane	Smith	2001-03-04	Nelson
adory	April	Dory	2001-03-04	Paige
shu	Simon	Hu	2003-01-29	Paige
jsmith2	Jane	Smith	2006-11-09	Nelson
lortiz	Lorena	Ortiz	2010-09-26	Nelson

titles

title	author	genre	word_count
The House on the Hill	jsmith	contemporary	106789

title	author	genre	word_count
The Blue Diary	jsmith	contemporary	95019
Dragon Eaters	adory	fantasy	135501
Silent Wizards	adory	fantasy	126038
Forbidden Alchemy	adory	fantasy	111666
My Father’s Piano	shu	memoir	101365
Blueberry Pastures	jsmith2	contemporary	95019
Sudden Confinement	jsmith2	horror	95134
Rubi Saves the World	lortiz	young adult	76045

sales

title	rights	advance	royalty
The House on the Hill	domestic first print	15000	0.125
Dragon Eaters	domestic first print	12000	0.100
Dragon Eaters	foreign markets	5000	0.050
Dragon Eaters	audio	4000	0.075
Blueberry Pastures	domestic first print	15000	0.125
My Father’s Piano	domestic first print	14500	0.100
My Father’s Piano	foreign markets	14500	0.100
Rubi Saves the World	domestic first print	13500	0.110
Rubi Saves the World	audio	6000	0.060

Your Answer:

```
SELECT agent, genre, count()

FROM clients c JOIN titles t

ON c.cid = t.author

GROUP BY t.genre

ORDER BY c.agent
```

```
SELECT agent, genre, COUNT()
          FROM titles
          JOIN clients
          ON author = cid
          GROUP BY agent, genre
          ORDER BY agent
```

Missing agent in group by

Question 30

10 / 10 pts

Provide a SQL-style pseudocode strategy (using relational data concepts such as SELECT, WHERE, GROUP BY, and JOIN) for solving the problem.

Find the average word count for books of each genre.

The following three data tables describe the authors/clients, book titles, and sales to publishers for a certain literary agency:

clients

cid	first_name	last_name	sign_date	agent
jsmith	Jane	Smith	2001-03-04	Nelson
adory	April	Dory	2001-03-04	Paige
shu	Simon	Hu	2003-01-29	Paige
jsmith2	Jane	Smith	2006-11-09	Nelson
lortiz	Lorena	Ortiz	2010-09-26	Nelson

titles

title	author	genre	word_count
The House on the Hill	jsmith	contemporary	106789
The Blue Diary	jsmith	contemporary	95019
Dragon Eaters	adory	fantasy	135501
Silent Wizards	adory	fantasy	126038

title	author	genre	word_count
Forbidden Alchemy	adory	fantasy	111666
My Father’s Piano	shu	memoir	101365
Blueberry Pastures	jsmith2	contemporary	95019
Sudden Confinement	jsmith2	horror	95134
Rubi Saves the World	lortiz	young adult	76045

sales

title	rights	advance	royalty
The House on the Hill	domestic first print	15000	0.125
Dragon Eaters	domestic first print	12000	0.100
Dragon Eaters	foreign markets	5000	0.050
Dragon Eaters	audio	4000	0.075
Blueberry Pastures	domestic first print	15000	0.125
My Father’s Piano	domestic first print	14500	0.100
My Father’s Piano	foreign markets	14500	0.100
Rubi Saves the World	domestic first print	13500	0.110
Rubi Saves the World	audio	6000	0.060

Your Answer:

SELECT genre, AVG(word_count) FROM titles GROUP BY genre

```
SELECT genre, AVG(word_count)
         FROM titles
         GROUP BY genre
```

Question 31

10 / 10 pts

Provide a SQL-style pseudocode strategy (using relational data concepts such as SELECT, WHERE, GROUP BY, and JOIN) for solving the problem.

Rank the genres from largest advances to smallest advances, using the average advance from domestic first print rights.

The following three data tables describe the authors/clients, book titles, and sales to publishers for a certain literary agency:

clients

cid	first_name	last_name	sign_date	agent
jsmith	Jane	Smith	2001-03-04	Nelson
adory	April	Dory	2001-03-04	Paige
shu	Simon	Hu	2003-01-29	Paige
jsmith2	Jane	Smith	2006-11-09	Nelson
lortiz	Lorena	Ortiz	2010-09-26	Nelson

titles

title	author	genre	word_count
The House on the Hill	jsmith	contemporary	106789
The Blue Diary	jsmith	contemporary	95019
Dragon Eaters	adory	fantasy	135501
Silent Wizards	adory	fantasy	126038
Forbidden Alchemy	adory	fantasy	111666
My Father’s Piano	shu	memoir	101365
Blueberry Pastures	jsmith2	contemporary	95019
Sudden Confinement	jsmith2	horror	95134
Rubi Saves the World	lortiz	young adult	76045

sales

title	rights	advance	royalty
The House on the Hill	domestic first print	15000	0.125
Dragon Eaters	domestic first print	12000	0.100
Dragon Eaters	foreign markets	5000	0.050
Dragon Eaters	audio	4000	0.075
Blueberry Pastures	domestic first print	15000	0.125
My Father’s Piano	domestic first print	14500	0.100
My Father’s Piano	foreign markets	14500	0.100
Rubi Saves the World	domestic first print	13500	0.110
Rubi Saves the World	audio	6000	0.060

Your Answer:

SELECT genre, AVG(s.advance) FROM sales s

INNER JOIN titles t ON s.title = t.title

WHERE s.rights = 'domestic first print'

GROUP BY t.genre

ORDER BY AVG(s.advance) DESC

```
SELECT genre, AVG(advance)
           FROM sales
           JOIN titles
           ON titles.title = sales.title
           WHERE rights = "domestic first print"
 GROUP BY genre
           ORDER BY AVG(advance) DESC
```

Quiz Score: **130** out of 155