# Data Engineering: Midterm Exam

IMEN891K / IDSC501 / AXGS-501, Fall 2026

**Name:** _____ **Student ID:** _____

## Part I: True/False (20 points)

Total: 40 points. Each item is worth 2 points. Mark one box per item.

**1.** A temperature sensor acts as an encoder: ☒ T ☐ F
it converts a physical signal into a form that can be recorded digitally.

**2.** Production volume and defect rate are strongly correlated. ☐ T ☒ F
This proves that lowering production will lower the defect rate.

**3.** Some customer records have missing values. ☐ T ☒ F
The team should fill them all before checking why they are missing.

**4.** Two systems use different customer IDs. ☐ T ☒ F
Giving their columns the same names is enough to match each customer.

**5.** A model measures customer similarity using age and income. ☒ T ☐ F
Scaling helps stop income's larger numbers from dominating the distance.

**6.** A model must distinguish low, medium and high priority. ☐ T ☒ F
One-hot encoding tells the model which level is higher.

**7.** A team wants fewer inputs while keeping each variable's meaning. ☐ T ☒ F
Feature selection does this by combining variables into new ones.

**8.** A relational database can store each customer's details once. ☒ T ☐ F
Orders can refer to that customer using a key.

**9.** Some customer profiles have fields that others do not. ☒ T ☐ F
A document store can keep these profiles in the same collection.

**10.** A SQL query's result can be a table in another query. ☒ T ☐ F
A MongoDB aggregation stage can pass its documents to the next stage.

## Part II: Multiple Choice (20 points)

Choose exactly one option per item. Each item is worth 2 points.

**11.** **Answer:** **B**

You need data from two sites: one has an API; the other only has web pages.
Which methods should you use for the first and second sites?

A. Web crawling; API calls
B. API calls; web crawling
C. API calls; sensor measurement
D. Synthetic data; web crawling

**Part II: Multiple Choice (continued)**

**12.** **Answer: D**

A few houses have prices far above the rest.
Which measure of a typical price is less affected by these extreme values?

A. Mean
B. Range
C. Standard deviation
D. Median

**13.** **Answer: C**

Two stores have the same means, variances and correlation for sales and ad spending.
Why should you still compare their scatter plots?

A. Matching statistics mean the data have the same shape
B. Matching statistics mean the data have the same outliers
C. Different patterns can produce the same statistics
D. A scatter plot cannot show anything beyond the correlation

**14.** **Answer: A**

Sensor readings fluctuate rapidly because of noise.
You want a smoother trend on a dashboard. Which method should you use?

A. Moving average
B. One-hot encoding
C. kNN imputation
D. Z-score standardization

**15.** **Answer: C**

You need a sample that includes every factory.
Some factories have far more records than others. What should you do?

A. Take the first records in the combined dataset
B. Sample only from the largest factory
C. Randomly sample at least one record from every factory
D. Remove unusually large or small values

For items 16–18, every product has a matching company:
`Product(name, category, price, company_id)`
`Company(company_id, country)`, with unique `company_id`.

```
SELECT name
FROM Product
WHERE category = 'Electronics';
```

**16.** **Answer: A**

This query builds a product list. What does the list contain?

A. Names of Electronics products
B. All fields of Electronics products
C. Names of all products
D. The number of Electronics products

**Part II: Multiple Choice (continued)**

Every product has a matching company.
`Product(name, category, price, company_id)`
`Company(company_id, country)`, with unique `company_id`.

```
SELECT p.name, c.country
FROM Product p JOIN Company c
  ON p.company_id = c.company_id;
```

**17.** **Answer: D**

This query builds a supplier report. What does the report show?

A. The number of products in each country

B. Every possible product–company pair

C. Each product and its category

D. Each product and the country of its company

```
SELECT category, AVG(price) AS average_price
FROM Product
GROUP BY category;
```

**18.** **Answer: C**

A dashboard runs this query. What does each result row describe?

A. Each product and its price

B. One average price for all products

C. Each category and its average product price

D. Each company and its average product price

**19.** **Answer: B**

A service must (i) find a session by its token and (ii) follow links between friends. Which NoSQL types fit these tasks, in that order?

A. Graph; key-value

B. Key-value; graph

C. Document; wide-column

D. Wide-column; document

**20.** **Answer: D**

In a document store, each sale records a city, product category and sale amount. You need total sales for Electronics in each city. Which steps should you use?

A. Keep Electronics sales; group by category; sum the amounts

B. Keep Electronics sales; group by city; count the sales

C. Group all sales by city; sum the amounts; sort the totals

D. Keep Electronics sales; group by city; sum the amounts