25F - Data Engineering (IMEN891K/IDSC501)

Midterm Exam (Total: 25 points)

Name: _____ **Student ID:** _____

Part I: True/False (11 points)

Answer each statement as True or False. Each question is worth 1 point.

1. Most of the difficulty in data science comes from modeling, with other steps like data collection considered relatively straightforward. ❑ T ✓ F
2. While most data today exist in digital form, offline sources such as sensor data remain crucial for representing real-world conditions accurately. ✓ T ❑ F
3. The ELT pipeline fits better with traditional systems, while the ETL pipeline is more suitable for modern cloud-based services. ❑ T ✓ F
4. In data integration, schema mapping aligns the structures of different data sources, while data fusion combines overlapping records into a unified view. ✓ T ❑ F
5. Both PCA and t-SNE are dimensionality reduction methods; PCA can reduce data to any number of dimensions and is also used for feature extraction, while t-SNE is mainly used for visualization in two dimensions. ✓ T ❑ F
6. Time-series data are decomposed according to their purpose; economic data are often divided into trend, cyclical, and seasonal components, while sensor data are decomposed into frequency components using Fourier analysis. ✓ T ❑ F
7. Regression of X on Y and regression of Y on X are mathematically identical, and the fitted models produce the same predicted values. ❑ T ✓ F
8. If there are missing values in a dataset, it is always better to apply data imputation methods to fill them in for reliable pre-processing. ❑ T ✓ F
9. One-hot encoding is suitable for ordinal categorical variables because it preserves the natural order between categories. ❑ T ✓ F
10. Positively skewed data can be normalized using the log transformation, while negatively skewed data can be normalized by its reflection. ✓ T ❑ F
11. Foundation models are useful for feature extraction; for example, DINOv3 extracts visual features, while CLIP captures image-text joint features. ✓ T ❑ F

Part II: Short Answer (6 points)

Answer each question using keyword terms. Each question is worth 2 points.

1. What branch of statistics focuses on developing methods that are less sensitive to outliers, such as the median or trimmed mean?

   Answer: ___robust___ statistics

2. How can we measure the correlation in non-linear relationships?

   Answer: rank correlation

3. What process identifies the key features to improve model performance and interpretability?

   Answer: feature selection

Part III: Code Completion (8 points)

Complete the missing parts of each SQL query. Each blank is worth 0.5 points.

1. The following schema defines two tables: `Product(pid, pname, manufacturer)` and `Company(cname, country)`. (1) Complete the SQL statement to declare a foreign key constraint, and (2) complete the query to list the product name and manufacturer country for companies in the USA.

```
-- (1) Declare the foreign key constraint
CREATE TABLE Product (
    pid INT ____PRIMARY____ KEY,
    pname VARCHAR(50),
    manufacturer VARCHAR(50),
    ____FOREIGN____ KEY ( ____manufacturer____ )
        REFERENCES ____Company____ ( ____cname____ )
);

-- (2) Query product-country pairs
____SELECT____ pname, country
____FROM____ Product
____JOIN____ Company
ON Product.____manufacturer____ = Company.____cname____
____WHERE____ country = 'USA';
```

2. Retrieve employees whose salary is above both their department's average and the overall company average, using the table `Employees(name, department, salary)`.

```
SELECT name
FROM Employees e
WHERE e.salary > (
    SELECT AVG(salary)
    FROM ____Employees____
    WHERE ____department____ = e.____department____
)
____AND____ e.salary > (
    SELECT AVG(salary)
    FROM ____Employees____
);
```