```
import logging
# Set up basic logging
logging.basicConfig(
    filename='etl_log.txt',
    level=logging.INFO,
    format='%(asctime)s - %(levelname)s - %(message)s'
)
import pandas as pd
import numpy as np
import mysql.connector
import datetime

csv_file_path = 'employees1.csv'
df = pd.read_csv(csv_file_path)
print("Raw data loaded:")
print(df.head())
print(df.columns.tolist())
logging.info("CSV loaded successfully.")

# Handle missing values
df.fillna({
    'EMAIL': 'not_provided@example.com',
    'PHONE_NUMBER': '0000000000',
```

```
    'HIRE_DATE': '01-Jan-00',
    'SALARY': 0
}, inplace=True)

# Standardize column names (optional)
df.columns = [col.strip().lower().replace(' ', '_') for col in df.columns]

print(df.columns.tolist())
# Convert hire_date from 'dd-MMM-yy' to 'YYYY-MM-DD'
df['hire_date'] = pd.to_datetime(df['hire_date'], format='%d-%b-%y',
errors='coerce')


# Replace invalid dates with a default
df['hire_date'] = df['hire_date'].fillna(pd.to_datetime('2000-01-01'))
# Replace non-numeric salaries with 0
df['salary'] = pd.to_numeric(df['salary'], errors='coerce').fillna(0).astype(int)

logging.info("Data cleaning completed.")

# ----------------------------
# Step 3: Load - Insert into MySQL
# ----------------------------

# MySQL connection
```

```
mydb = mysql.connector.connect(
  host="localhost",
  user="root",
  password="root",
  database="employee"  # <-- change this
)

cursor = mydb.cursor()

# Create table if not exists
cursor.execute("""
  CREATE TABLE IF NOT EXISTS salary_2 (
    empid INT,
    firstname VARCHAR(50),
    lastname VARCHAR(50),
    email VARCHAR(100),
    phone VARCHAR(20),
    hire_date DATE,
    job_id VARCHAR(20),
    salary INT
  )
""")

# Insert each row into the table
for index, row in df.iterrows():
```

```
    sql = """
        INSERT INTO salary_2 (empid, firstname, lastname, email, phone,
hire_date, job_id, salary)
        VALUES (%s, %s, %s, %s, %s, %s, %s, %s)
        ON DUPLICATE KEY UPDATE
          firstname=VALUES(firstname),
          lastname=VALUES(lastname),
          email=VALUES(email),
          phone=VALUES(phone),
          hire_date=VALUES(hire_date),
          job_id=VALUES(job_id),
          salary=VALUES(salary)
    """
    values = (
        int(row['employee_id']),
        row['first_name'],
        row['last_name'],
        row['email'],
        row['phone_number'],
        row['hire_date'].date(),
        row['job_id'],
        int(row['salary'])
    )
    cursor.execute(sql, values)
```

```
mydb.commit()

cursor.close()

mydb.close()

logging.error("Something went wrong", exc_info=True)

print("ETL process completed successfully.")
```

```
import os

print(os.getcwd())
```

```
get_ipython().system('jupyter nbconvert --to script ETL.ipynb')
```