Lab 17 AI for Data Processing: Data Cleaning and Preprocessing Scripts

Name: B.yashwanth

Enrollment Number: 2503A51L42

Task 1 :

Task : Use AI to generate a Python script for cleaning an employee dataset

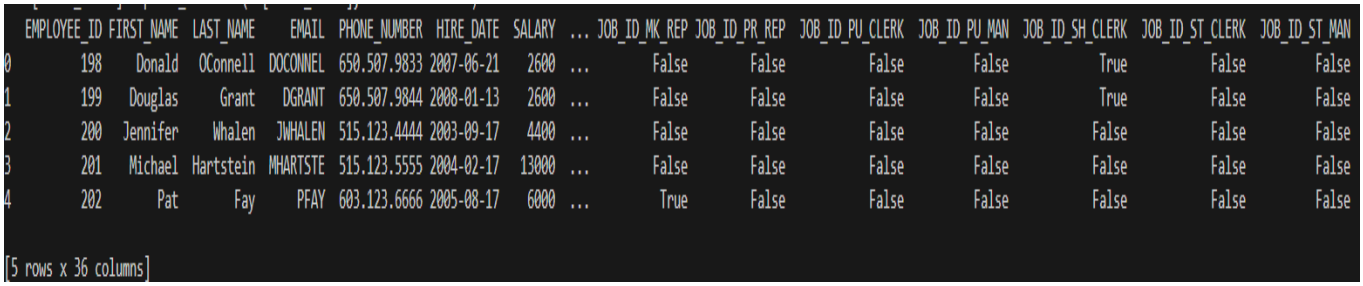
Prompt: Generate a Python script to clean an employee dataset by handling missing values, formatting joining_date, standardizing department names, and encoding categorical variables.

Code:

employee_data_cleaning > clean_employee_data.py

```python
import pandas as pd
import os

# Print current working directory
print("Current working directory:", os.getcwd())

# Load dataset
df = pd.read_csv(r"C:\Users\Suhana Rehan\OneDrive\Desktop\AI assistant coding\employee_data_cleaning\employees.csv")

# Handle missing values (safe assignment)
df = df.fillna({
    'SALARY': df['SALARY'].median(),
    'DEPARTMENT_ID': 'Unknown',
    'HIRE_DATE': '1900-01-01'
})

# Convert HIRE_DATE to datetime with format hint (optional but safer)
df['HIRE_DATE'] = pd.to_datetime(df['HIRE_DATE'], errors='coerce')

# Standardize DEPARTMENT_ID
df['DEPARTMENT_ID'] = df['DEPARTMENT_ID'].astype(str).str.lower().replace({
    'human resources': 'hr',
    'hr': 'hr',
    'finance': 'finance',
    'sales': 'sales',
    'marketing': 'marketing'
})


# Encode categorical variables
df_encoded = pd.get_dummies(df, columns=['DEPARTMENT_ID', 'JOB_ID'])

# Save cleaned data
df_encoded.to_csv("cleaned_employees.csv", index=False)

# Display cleaned DataFrame
print(df_encoded.head())
```

Output:



	EMPLOYEE_ID	FIRST_NAME	LAST_NAME	EMAIL	PHONE_NUMBER	HIRE_DATE	SALARY	...	JOB_ID_MK_REP	JOB_ID_PR_REP	JOB_ID_PU_CLERK	JOB_ID_PU_MAN	JOB_ID_SH_CLERK	JOB_ID_ST_CLERK	JOB_ID_ST_MAN
0	198	Donald	OConnell	DOCONNEL	650.507.9833	2007-06-21	2600	...	False	False	False	False	True	False	False
1	199	Douglas	Grant	DGRANT	650.507.9844	2008-01-13	2600	...	False	False	False	False	True	False	False
2	200	Jennifer	Whalen	JWHALEN	515.123.4444	2003-09-17	4400	...	False	False	False	False	False	False	False
3	201	Michael	Hartstein	MHARTSTE	515.123.5555	2004-02-17	13000	...	False	False	False	False	False	False	False
4	202	Pat	Fay	PFAY	603.123.6666	2005-08-17	6000	...	True	False	False	False	False	False	False

[5 rows x 36 columns]

Observations:

The AI help me clean the the database and it handled all the missing value and made the database ready to use

Task 2 :

Task:

Use AI to generate a script for preprocessing a sales transaction dataset

Prompt:

Preprocess a sales dataset by parsing dates, extracting Month-Year, removing invalid amounts, and normalizing values.

Code:

sales_data_cleaning > clean_sales_data.py

```python
import pandas as pd
from sklearn.preprocessing import MinMaxScaler

# Load dataset
df = pd.read_csv(r"""C:\Users\Suhana Rehan\OneDrive\Desktop\AI assistant coding\sales_data_cleaning\transactions.csv"

# Convert transaction_date to datetime
df['transaction_date'] = pd.to_datetime(df['transaction_date'], format='%Y-%m-%d', errors='coerce')

# Create Month-Year column
df['Month_Year'] = df['transaction_date'].dt.to_period('M').astype(str)

# Remove rows with zero or negative transaction_amount
df = df[df['transaction_amount'] > 0]

# Normalize transaction_amount using Min-Max scaling
scaler = MinMaxScaler()
df['normalized_amount'] = scaler.fit_transform(df[['transaction_amount']])

# Save cleaned data
df.to_csv("cleaned_transactions.csv", index=False)

# Display first few rows
print(df.head())
with open(r"C:\Users\Suhana Rehan\OneDrive\Desktop\AI assistant coding\sales_data_cleaning\transactions.csv", "r") as
    print(f.read())

```

Output:

```
   transaction_id customer_id transaction_date  transaction_amount product_category Month_Year  normalized_amount
0               1        C001       2025-01-15                 250      Electronics    2025-01           0.095238
3               4        C004       2025-02-18                1200      Electronics    2025-02           1.000000
4               5        C005       2025-03-10                 300          Fashion    2025-03           0.142857
5               6        C006       2025-03-15                 450          Grocery    2025-03           0.285714
7               8        C008       2025-04-12                 800          Fashion    2025-04           0.619048
transaction_id,customer_id,transaction_date,transaction_amount,product_category
1,C001,2025-01-15,250,Electronics
2,C002,2025-01-20,0,Fashion
3,C003,2025-02-05,-50,Grocery
4,C004,2025-02-18,1200,Electronics
5,C005,2025-03-10,300,Fashion
6,C006,2025-03-15,450,Grocery
7,C007,2025-04-01,0,Electronics
8,C008,2025-04-12,800,Fashion
9,C009,2025-05-05,150,Grocery
10,C010,2025-05-20,600,Electronics
```

Observation:

- The AI helped me in cleaning the data, handle the missing value and irrelevant information in my database
- It made my database readable and ready to process

Task 3

Task:

Use AI to generate a script for cleaning healthcare patient records.

Prompt:

Clean healthcare patient records by imputing numeric means, standardizing height units, fixing gender labels, and dropping IDs.

Code:

Output:

Observation:

Task 4

Use AI to write a script to preprocess a social media text dataset.

Prompt:

Code:

Output:

Observation:

Task 5

Task:

Use AI to create a preprocessing script for a financial dataset

Prompt:

Code:

Output:

Observation: