Python Data Preprocessing Tasks

1. Handle Missing Data:

num.py > ...

```python
import pandas as pd
data = {'A': [1, 2, 2], 'B': [4, 5, 5]}
df = pd.DataFrame(data)

df_unique = df.drop_duplicates()

print("DataFrame after removing duplicates:")
print(df_unique)

```

2. Remove Duplicate Rows:

Code:

num.py > ...

```python
import pandas as pd
data = {'A': [1, 2, 2], 'B': [4, 5, 5]}
df = pd.DataFrame(data)

df_unique = df.drop_duplicates()

print("DataFrame after removing duplicates:")
print(df_unique)

```

3. Detect and Remove Outliers (Z-Score Method):

Code:

num.py > ...

```python
from scipy.stats import zscore
import pandas as pd

data = {'A': [10, 12, 11, 100], 'B': [5, 6, 5, 50]}
df = pd.DataFrame(data)


z_scores = zscore(df)
df_no_outliers = df[(abs(z_scores) < 3).all(axis=1)]

print("DataFrame after removing outliers:")
print(df_no_outliers)

```

4. Convert Categorical Data to Numerical (One-Hot Encoding):

Code:

num.py > ...

```python
import pandas as pd
data = {'Category': ['A', 'B', 'A']}
df = pd.DataFrame(data)


df_encoded = pd.get_dummies(df, columns=['Category'])

print("DataFrame after one-hot encoding:")
print(df_encoded)

```

5. Normalize Data Using Min-Max Normalization:

Code:

num.py > ...

```python
import pandas as pd
data = {'A': [10, 20, 30], 'B': [100, 200, 300]}
df = pd.DataFrame(data)

df_normalized = (df - df.min()) / (df.max() - df.min())

print("DataFrame after min-max normalization:")
print(df_normalized)

```

6. Clean Text Data:

Code:

num.py > ...

```python
import re

text = "Hello, World! Welcome to Data Science 101."

cleaned_text = re.sub(r'[^a-zA-Z0-9\s]', '', text).lower()

print("Cleaned text:")
print(cleaned_text)

```

What I Learned:

Through these exercises, I deepened my understanding of essential data preprocessing techniques. Handling missing data using the mean improves dataset quality for analysis. Removing duplicates ensures data consistency. Detecting and eliminating outliers with Z-scores maintains data integrity for statistical

analysis. Transforming categorical variables with one-hot encoding allows machine learning models to process non-numerical data. Min-max normalization standardizes data ranges, enhancing comparability.

Finally, text cleaning prepares raw text for natural language processing by removing noise and standardizing case. These techniques are fundamental for building reliable and accurate machine learning models.