# Scaling Techniques Documentation

## 1. Handling Missing Data

To ensure dataset integrity and avoid errors during modeling:
- Numeric features → Missing values were imputed using the median (robust to outliers).
- Categorical features → Missing values were imputed using the most frequent value (mode).

```
from sklearn.impute import SimpleImputer

# Numeric imputation
num_cols = df.select_dtypes(include=['int64','float64']).columns
num_imputer = SimpleImputer(strategy="median")
df[num_cols] = num_imputer.fit_transform(df[num_cols])

# Categorical imputation
cat_cols = df.select_dtypes(include=['object']).columns
cat_imputer = SimpleImputer(strategy="most_frequent")
df[cat_cols] = cat_imputer.fit_transform(df[cat_cols])
```

## 2. Encoding Categorical Variables

Categorical features were transformed into numeric format using One-Hot Encoding.
- drop='first' → avoids dummy variable trap.
- sparse_output=False → ensures dense arrays for Pandas DataFrames.
- handle_unknown='ignore' → prevents errors with unseen categories.

```
from sklearn.preprocessing import OneHotEncoder
import pandas as pd

encoder = OneHotEncoder(drop="first", sparse_output=False, handle_unknown="ignore")
encoded = encoder.fit_transform(df[cat_cols])

# Get encoded feature names
encoded_cols = encoder.get_feature_names_out(cat_cols)

# Create encoded DataFrame
encoded_df = pd.DataFrame(encoded, columns=encoded_cols)

# Merge encoded data with rest of dataset
df_encoded = pd.concat([df.drop(columns=cat_cols), encoded_df], axis=1)
```

## 3. Feature Scaling

For model compatibility and performance, numerical features were standardized using StandardScaler. Categorical dummy variables (0/1) were left unchanged.

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
df[num_cols] = scaler.fit_transform(df[num_cols])
```

## 4. Outputs

- 01_missing_data_handled.csv → dataset with missing values handled.
- 02_categorical_encoded.csv → dataset after categorical encoding.
- Scaling_Techniques_Documentation.pdf → this documentation file.