Data Preprocessing Techniques Report

1- Introduction

This document details the preprocessing steps applied to the dataset in preparation for further analysis, specifically focusing on customer segmentation and churn analysis. The preprocessing phase is crucial for ensuring that the data fed into the analysis models is clean, appropriately formatted, and normalized.

2- Feature Encoding

Categorical Encoding: Many machine learning models require numerical input. Categorical variables were encoded using the Label Encoding

Label encoding assigns a unique integer to each category. While efficient in terms of space (since it doesn't add new columns to the dataset), it implies an ordinal relationship between categories which may not exist. This can be problematic for algorithms that assume numerical relationships between values.

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder



# Define a dictionary to map categorical values to numerical values
mapping = {
    'Contract': {'Month-to-month': 0, 'One year': 1, 'Two year': 2},
    'InternetService': {'DSL': 0, 'Fiber optic': 1, 'No': 2},
    'Churn': {'No': 0, 'Yes': 1},
    'MultipleLines': {'No phone service': 0, 'No': 1, 'Yes': 2},
    'PhoneService': {'No': 0, 'Yes': 1},
    'Dependents': {'No': 0, 'Yes': 1},
    'gender': {'Female': 0, 'Male': 1}
}

# Use the 'replace' method to map the categorical values
for column, value_map in mapping.items():
  data[column] = data[column].replace(value_map)

data.head()
```

	gender	SeniorCitizen	Dependents	tenure	PhoneService	MultipleLines	InternetService	Contract	MonthlyCharges	Churn
0	0	0	0	1	0	1	0	0	29.85	0
1	1	0	0	34	1	1	0	1	56.95	0
2	1	0	0	2	1	1	0	0	53.85	1
3	1	0	0	45	0	1	0	1	42.30	0
4	0	0	0	2	1	1	1	0	70.70	1

3- Feature Scaling

Standard Scaler: The Standard Scaler was applied to normalize the features of the dataset. This scaler removes the mean and scales the data to unit variance, which is particularly important for clustering algorithms that are sensitive to the scale of the data such as K-Means.

Reasons for Choosing Standard Scaler:

Unit Variance: Ensures that each feature contributes equally.

Normalization: Brings all features to the same scale, thus preventing any single feature from dominating the distance calculations in clustering algorithms.

Impact on Dataset's Attributes:

The scaling process adjusted the distribution of the data, ensuring that features with larger scales do not unduly influence the model's outcome.

It enhanced the algorithm's efficiency by standardizing the range of independent variables.

```
scaler = StandardScaler()
features_scaled = scaler.fit_transform(X)
```

Conclusion

The preprocessing steps outlined in this document are essential for ensuring that the data used in clustering and other machine learning tasks is clean, normalized, and appropriately encoded. By systematically cleaning, encoding, and scaling the data, we enhance the reliability of our analyses and ensure that any insights derived are based on the true patterns present in the data, not skewed by anomalies or scale differences.