In [1]:

```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression, LassoCV
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
from sklearn.model_selection import ShuffleSplit
from yellowbrick.regressor import PredictionError
from yellowbrick.model_selection import LearningCurve
from sklearn.decomposition import PCA
import seaborn as sns
from sklearn.metrics import mean_absolute_error

np.random.seed(1)

class DataScienceProject:
    def __init__(self):
        pass

    def load_data(self, file_path):
        # Load data using pandas
        data = pd.read_csv(file_path)
        return data

    def report_missing_values(self, df ):
        # Calculate the number of missing values per column
        missing_values = df.isnull().sum()
        missing_report = pd.DataFrame(missing_values, columns=['missing_values'])
        missing_report = missing_report[missing_report['missing_values'] > 0]

        # Suggest imputation values
        imputation_values = {}
        for column in missing_report.index:
            if df[column].dtype in ['int64', 'float64']:
                skewness = df[column].skew()
                if abs(skewness) > 0.5:
                    imputation_value = df[column].median()
                    imputation_values[column] = ('median', imputation_value)
                else:
                    imputation_value = df[column].mean()
                    imputation_values[column] = ('mean', imputation_value)
            else:
                imputation_value = df[column].mode()[0]
```

```
                imputation_values[column] = ('mode', imputation_value)

        return imputation_values

    def apply_imputations(self, df, imputation_values):
        for column, (strategy, value) in imputation_values.items():
            if strategy in ['mean', 'median', 'mode']:
                df[column].fillna(value, inplace=True)
        # print("after imp", df.isna().sum())
        return df


    def visualize_data(self, data, title_suffix=''):
        """
        Visualizes distributions of numerical and categorical features in the dataset.

        Args:
        data (DataFrame): The dataset to visualize.
        title_suffix (str): A suffix for the plot title to distinguish between original and preprocessed data.
        """
        numerical_cols = data.select_dtypes(include=['int64', 'float64']).columns
        categorical_cols = data.select_dtypes(include=['object']).columns

        # Plot for numerical features
        for col in numerical_cols:
            plt.figure(figsize=(8, 4))
            sns.histplot(data[col], kde=True)
            plt.title(f'Distribution of {col} {title_suffix}')
            plt.xlabel(col)
            plt.ylabel('Frequency')
            plt.show()

        # Plot for categorical features
        for col in categorical_cols:
            plt.figure(figsize=(8, 4))
            sns.countplot(x=col, data=data)
            plt.title(f'Distribution of {col} {title_suffix}')
            plt.xlabel(col)
            plt.ylabel('Count')
            plt.show()

    def select_data_within_iqr(self, df, iqr_factor=1.5):
        # Select only numerical columns for IQR calculation
        numerical_cols = df.select_dtypes(include=['int64', 'float64']).columns
```

```
        # Calculate IQR for numerical columns
        Q1 = df[numerical_cols].quantile(0.25)
        Q3 = df[numerical_cols].quantile(0.75)
        IQR = Q3 - Q1

        # Determine bounds for outlier detection
        lower_bound = Q1 - (iqr_factor * IQR)
        upper_bound = Q3 + (iqr_factor * IQR)

        # Create a filter for rows to keep
        filter_rows = ((df[numerical_cols] >= lower_bound) & (df[numerical_cols] <= upper_bound)).all(axis=1)

        # Apply this filter to the entire DataFrame
        filtered_df = df[filter_rows]

        return filtered_df

    def apply_log_transformation(self, df, target_column,skew_threshold=0.5):
        """
        Applies log transformation to highly skewed columns.

        Args:
        df (DataFrame): The dataframe containing the data.
        skew_threshold (float): The threshold to identify highly skewed columns.

        Returns:
        DataFrame: The dataframe with log-transformed columns.
        """
        for column in df.select_dtypes(include=['float64', 'int64']):
            if df[column].skew() > skew_threshold and column != target_column :
                df['Log_' + column] = np.log1p(df[column])

        # print(df.columns)
        # print("after log", df.isna().sum())
        return df

    def encode_categorical_columns(self, df):
        """
        Encodes categorical columns using one-hot encoding and removes the original columns.

        Args:
        df (DataFrame): The dataframe to process.

        Returns:
        DataFrame: The dataframe with categorical columns one-hot encoded.
```

```
        """
        categorical_cols = df.select_dtypes(include=['object']).columns

        for col in categorical_cols:
            # Apply one-hot encoding to each categorical column
            dummies = pd.get_dummies(df[col], prefix=col)
            df = pd.concat([df, dummies], axis=1)

            # Drop the original categorical column
            df.drop(col, axis=1, inplace=True)

        return df

    def preprocess_data(self, data, target_column):
        # Report and apply imputations, and handle outliers
        imputation_values = self.report_missing_values(data)
        data = self.apply_imputations(data, imputation_values)
        data = self.encode_categorical_columns(data)
        data = self.apply_log_transformation(data, target_column)
        data = self.select_data_within_iqr(data)
        return data

    def train_model(self, training_data, target_column):
        # Modify to return X_train, X_test, y_train, y_test for Lasso analysis
        X = training_data.drop(target_column, axis=1)
        y = training_data[target_column]
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

        model = LinearRegression()
        model.fit(X_train, y_train)
        return model, X_train, X_test, y_train, y_test

    def plot_actual_vs_predicted(self, y_test, y_predict, model_type):
        """
        Plots the actual vs predicted values.

        Args:
        y_test (array-like): The true values of the target variable.
        y_predict (array-like): The predicted values by the model.
        model_type (str): Type of the model ('Baseline' or 'Main').
        """
        plt.figure(figsize=(16, 6))
        plt.title(f"{model_type} Model: Actual vs Predicted Values")
        x_points = list(range(len(y_test)))
        plt.plot(x_points, y_test, label='Actual Values', marker='o')
```

```python
        plt.plot(x_points, y_predict, label='Predicted Values', marker='x')
        plt.xlabel('Data Points')
        plt.ylabel('Target Variable')
        plt.legend()
        plt.show()


    def plot_learning_curve_and_prediction_error(self, model, X_train, X_test, y_train, y_test, model_type):
        """
        Plots the learning curve and prediction error using Yellowbrick.

        Args:
        model: The trained model.
        X_train, X_test, y_train, y_test: Training and testing data.
        model_type (str): Type of the model ('Baseline' or 'Main').
        """
        # Learning Curve
        plt.figure(figsize=(10, 6))
        lc_viz = LearningCurve(model, cv=5, scoring='r2', n_jobs=4, train_sizes=np.linspace(0.1, 1.0, 10))
        lc_viz.fit(X_train, y_train)
        lc_viz.set_title(f"{model_type} Model: Learning Curve")
        lc_viz.show()

        # Prediction Error
        plt.figure(figsize=(10, 6))
        pe_viz = PredictionError(model)
        pe_viz.fit(X_train, y_train)
        pe_viz.score(X_test, y_test)
        pe_viz.set_title(f"{model_type} Model: Prediction Error")
        pe_viz.show()


    def feature_importance_lasso(self, X_train, y_train, X_test, y_test):
        # Create and fit the LassoCV model
        lasso = LassoCV(cv=5, random_state=0)
        lasso.fit(X_train, y_train)

        best_alpha = lasso.alpha_
        lasso_coef = lasso.coef_
        y_pred = lasso.predict(X_test)
        mse = mean_squared_error(y_test, y_pred)
        r2 = r2_score(y_test, y_pred)

        # Plotting feature importances
        plt.figure(figsize=(10, 6))
```

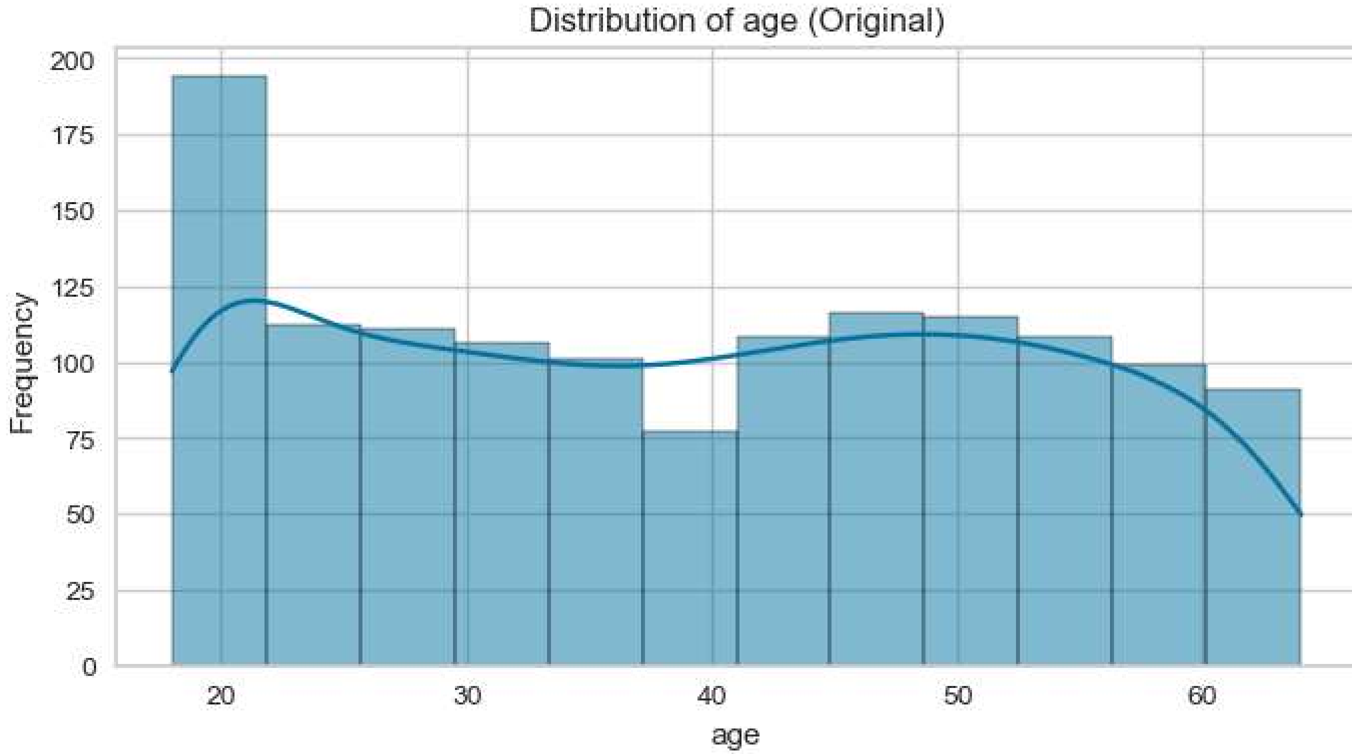
```
        feature_importance = pd.Series(lasso_coef, index=X_train.columns).sort_values()
        feature_importance.plot(kind='barh')
        plt.title('Feature Importances from Lasso Model')
        plt.xlabel('Importance')
        plt.ylabel('Features')
        plt.show()

        print(f"Best alpha: {best_alpha}")
        print(f"MSE: {mse}")
        print(f"R-squared: {r2:.2f}")

    def make_prediction(self, model, new_data):
        # Make predictions using the trained model
        prediction = model.predict(new_data)
        return prediction

    def train_and_evaluate(self, X_train, X_test, y_train, y_test):
        """
        Train a linear regression model and evaluate it.
        """
        model = LinearRegression()
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)

        mse = mean_squared_error(y_test, y_pred)
        r2 = r2_score(y_test, y_pred)
        mae = mean_absolute_error(y_test, y_pred)
        return model, mse, r2, mae

    def main_pipeline(self, file_path, target_column):
        print("===== Data Science Project Pipeline =====")
        print("[Step 1] Loading and Preprocessing Data")
        data = self.load_data(file_path)

        # Visualize original data
        self.visualize_data(data, title_suffix='(Original)')

        preprocessed_data = self.preprocess_data(data, target_column)

        # Visualize preprocessed data
        self.visualize_data(preprocessed_data, title_suffix='(Preprocessed)')

        # Baseline Model
        print("\n[Step 2] Training and Evaluating Baseline Model")
        mse_baseline, r2_baseline, baseline_model, X_train_baseline, X_test_baseline, y_train_baseline, y_test_baseline
```

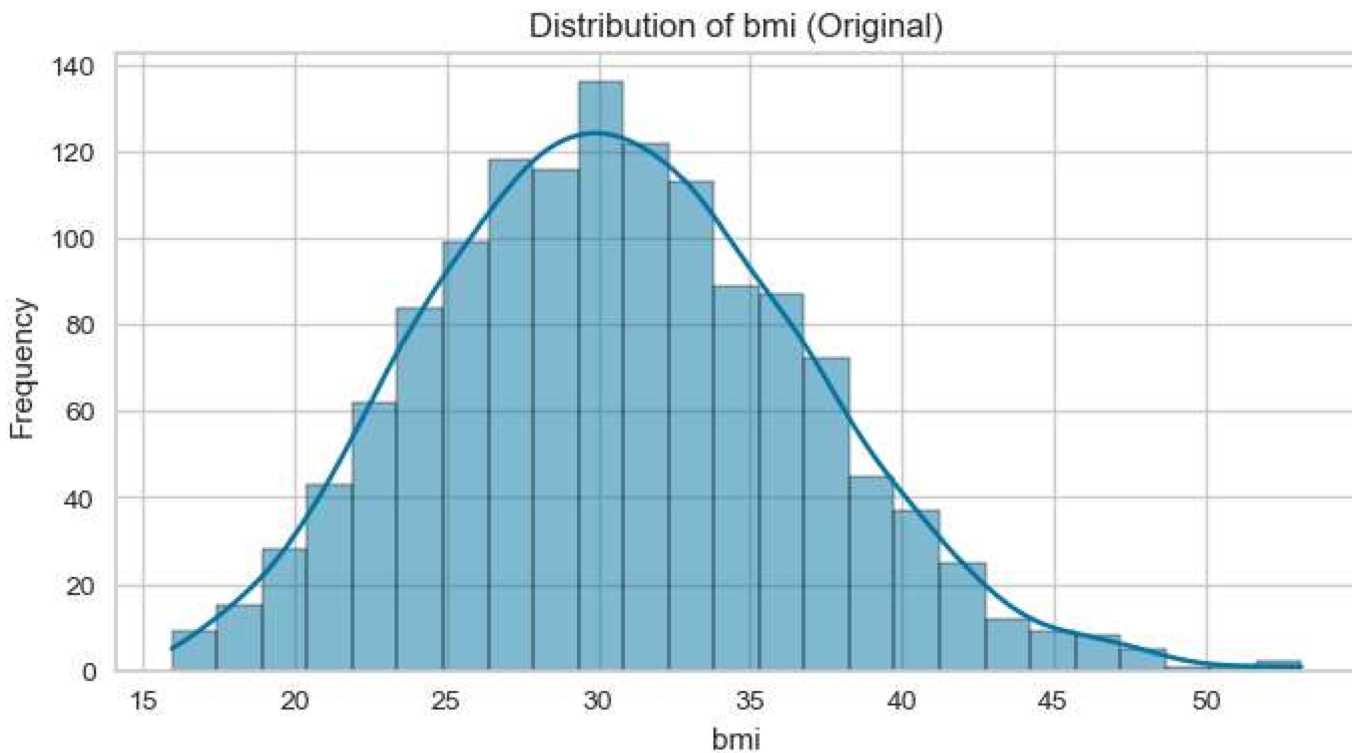
```
        y_pred_baseline = baseline_model.predict(X_test_baseline)
        self.plot_actual_vs_predicted(y_test_baseline, y_pred_baseline, "Baseline")
        self.plot_learning_curve_and_prediction_error(baseline_model, X_train_baseline, X_test_baseline, y_train_baseli

        # Main Model
        print("\n[Step 3] Training and Evaluating Main Model")
        model, X_train, X_test, y_train, y_test = self.train_model(preprocessed_data, target_column)
        _, mse_main, r2_main, mae = self.train_and_evaluate(X_train, X_test, y_train, y_test)
        y_predict_main = model.predict(X_test)
        self.plot_actual_vs_predicted(y_test, y_predict_main, "Main")
        self.plot_learning_curve_and_prediction_error(model, X_train, X_test, y_train, y_test, "Main")

        print("\n[Step 4] Performing Lasso Feature Importance Analysis")
        self.feature_importance_lasso(X_train, y_train, X_test, y_test)

        print("\n===== Model Comparison Results =====")
        print("Baseline Model:")
        print("  - MSE: {:.3f}".format(mse_baseline))
        print("  - R-squared: {:.3f}".format(r2_baseline))
        print("  - MAE: {:.3f}".format(mae))
        print("Main Model:")
        print("  - MSE: {:.3f}".format(mse_main))
        print("  - R-squared: {:.3f}".format(r2_main))
        print("  - MAE: {:.3f}".format(mae_baseline))
        print("====================================")

        return preprocessed_data

    def compare_with_baseline(self, data, target_column):
        # Updated to return the model and train/test splits
        baseline_data = data.dropna()
        X_baseline = baseline_data.drop(target_column, axis=1)
        X_baseline = self.encode_categorical_columns(X_baseline)
        y_baseline = baseline_data[target_column]
        X_train_baseline, X_test_baseline, y_train_baseline, y_test_baseline = train_test_split(X_baseline, y_baseline,
        baseline_model, mse_baseline, r2_baseline, mae_baseline = self.train_and_evaluate(X_train_baseline, X_test_base
        return mse_baseline, r2_baseline, baseline_model, X_train_baseline, X_test_baseline, y_train_baseline, y_test_b
```

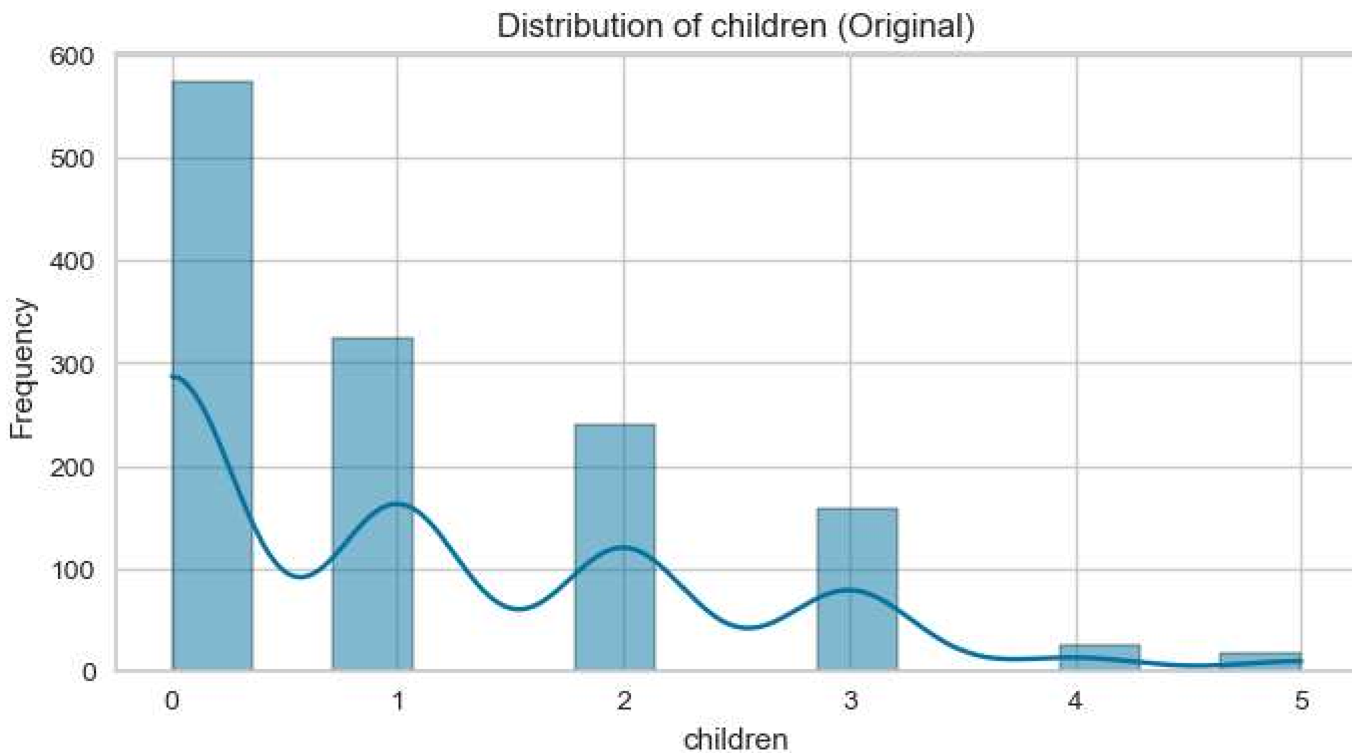
In [6]:

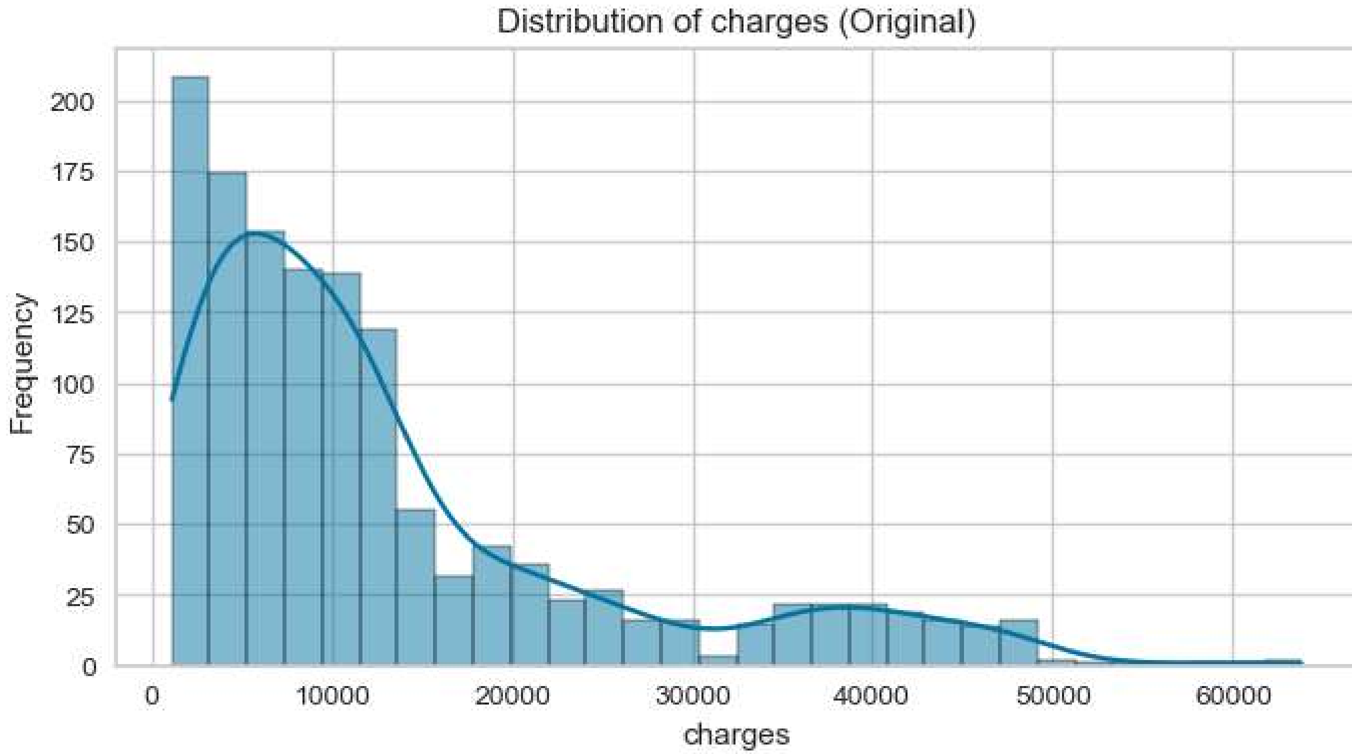
```
# Create an instance of the DataScienceProject class
dsp = DataScienceProject()

# Run the main pipeline
prediction = dsp.main_pipeline('D:\Temp\insurance.csv', 'charges')
```

```
===== Data Science Project Pipeline =====
[Step 1] Loading and Preprocessing Data
```

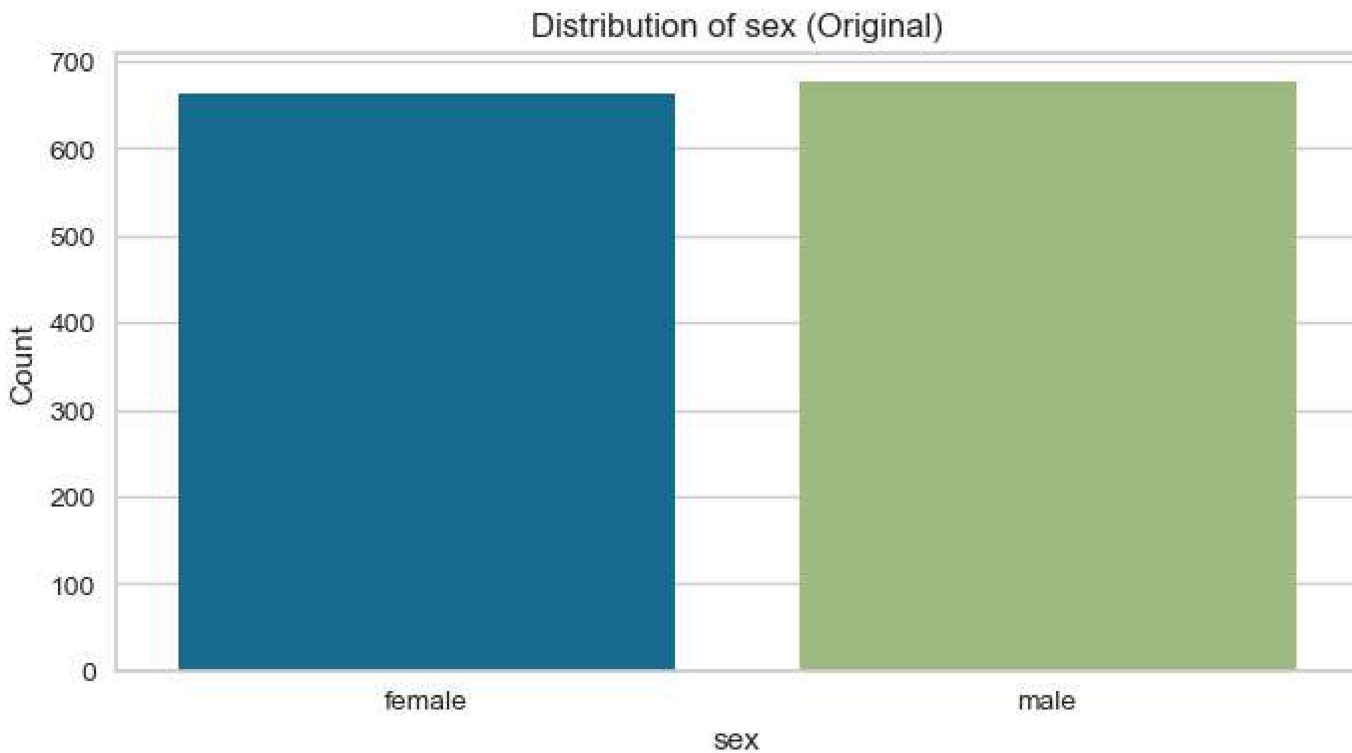


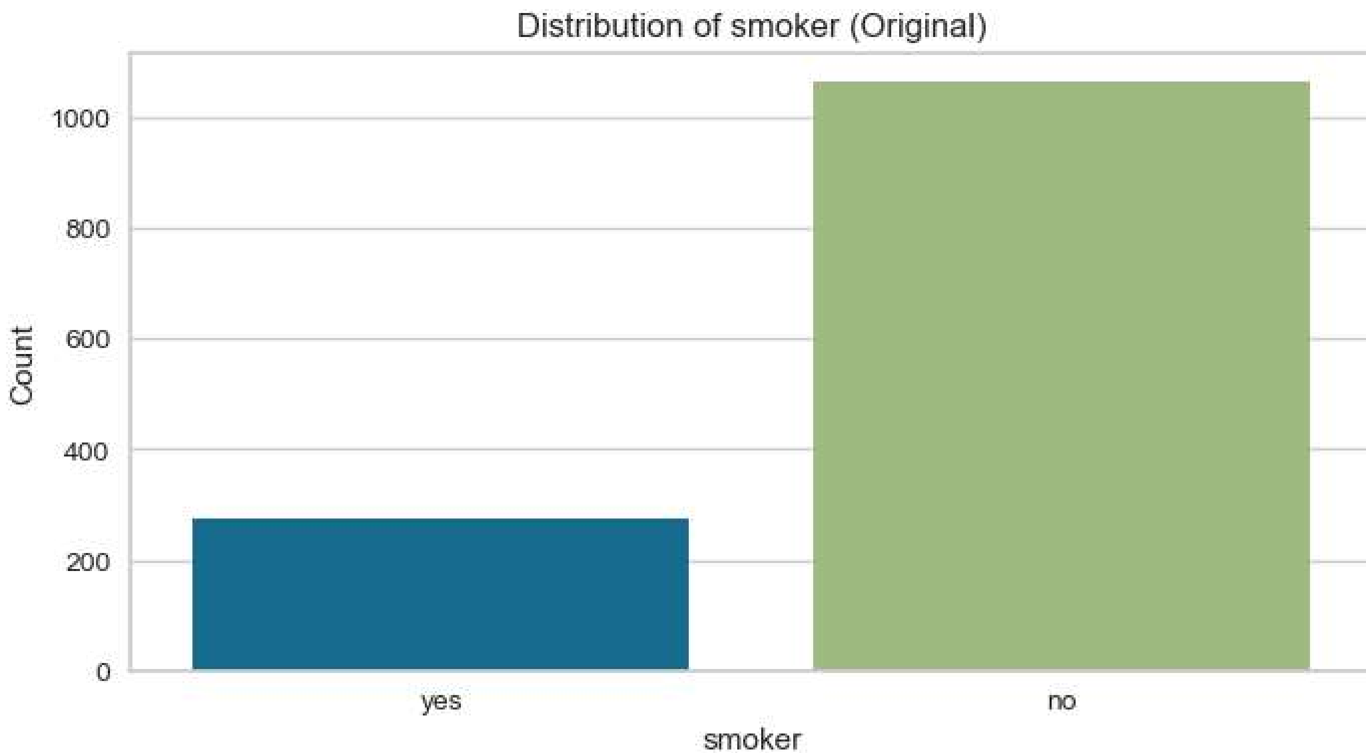

Distribution of bmi (Original)
Frequency
bmi
0
20
40
60
80
100
120
140
15
20
25
30
35
40
45
50

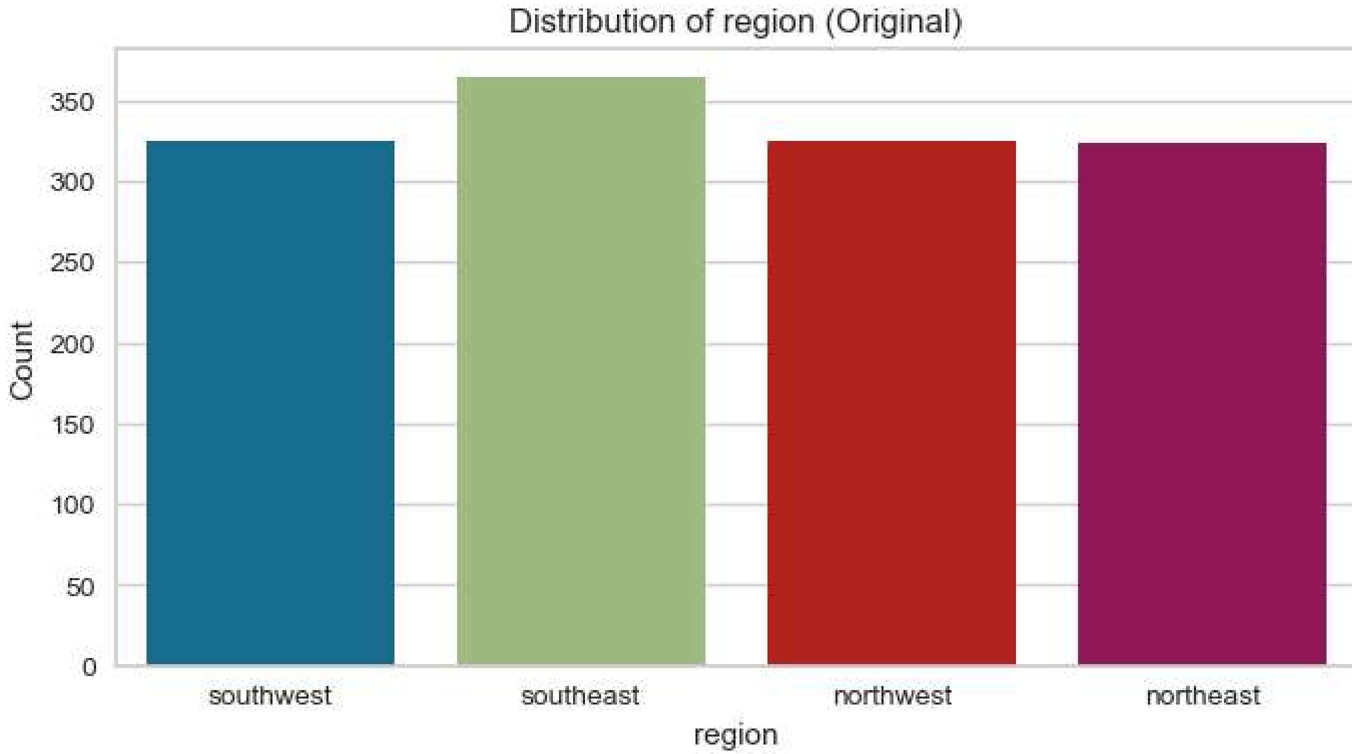

Distribution of children (Original)
Frequency
600
500
400
300
200
100
0
0
1
2
3
4
5
children

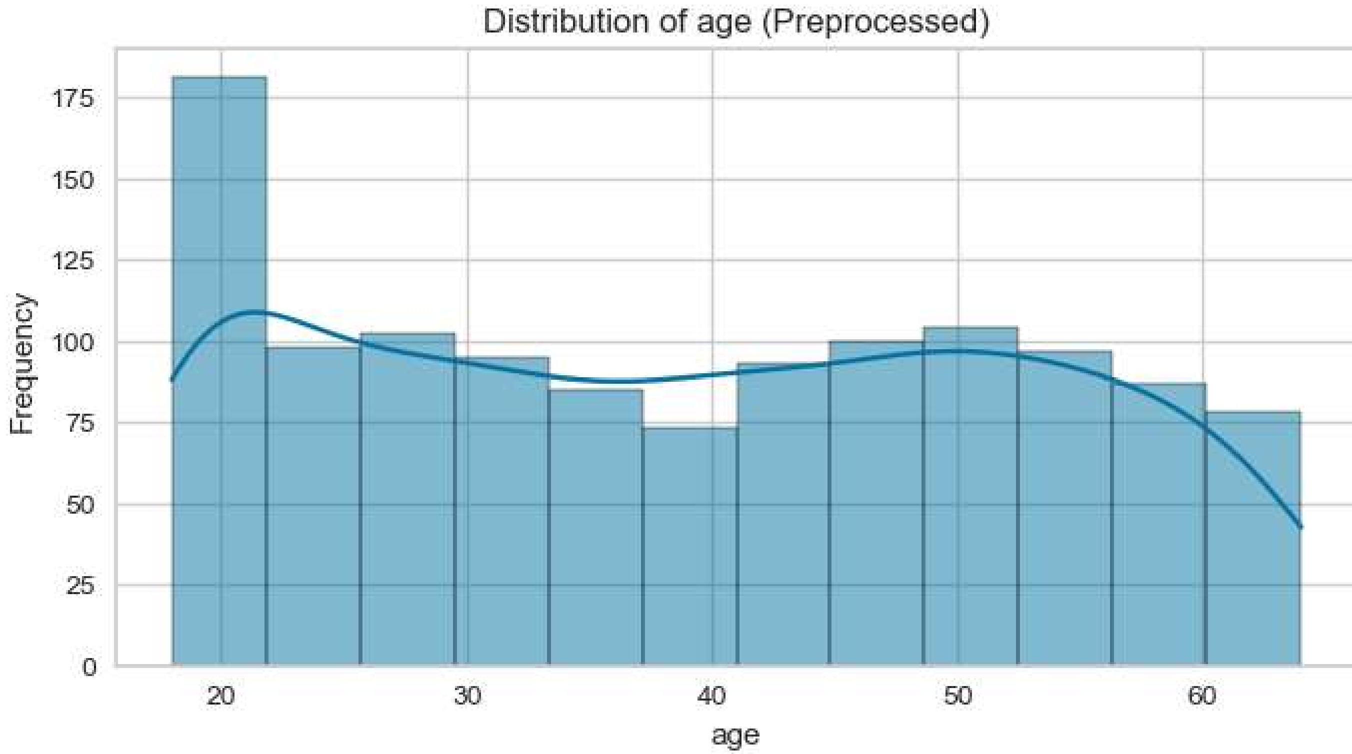


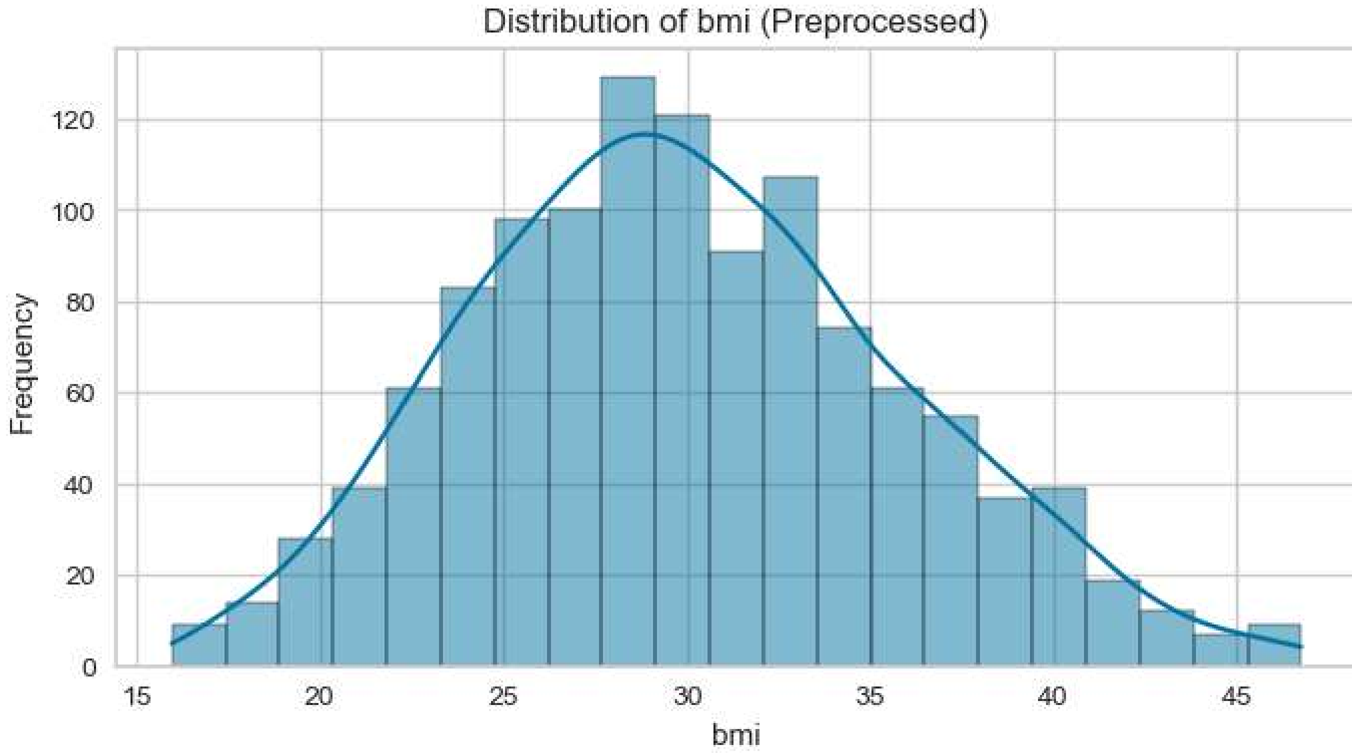
Distribution of charges (Original)
Frequency
charges
0
25
50
75
100
125
150
175
200
0
10000
20000
30000
40000
50000
60000

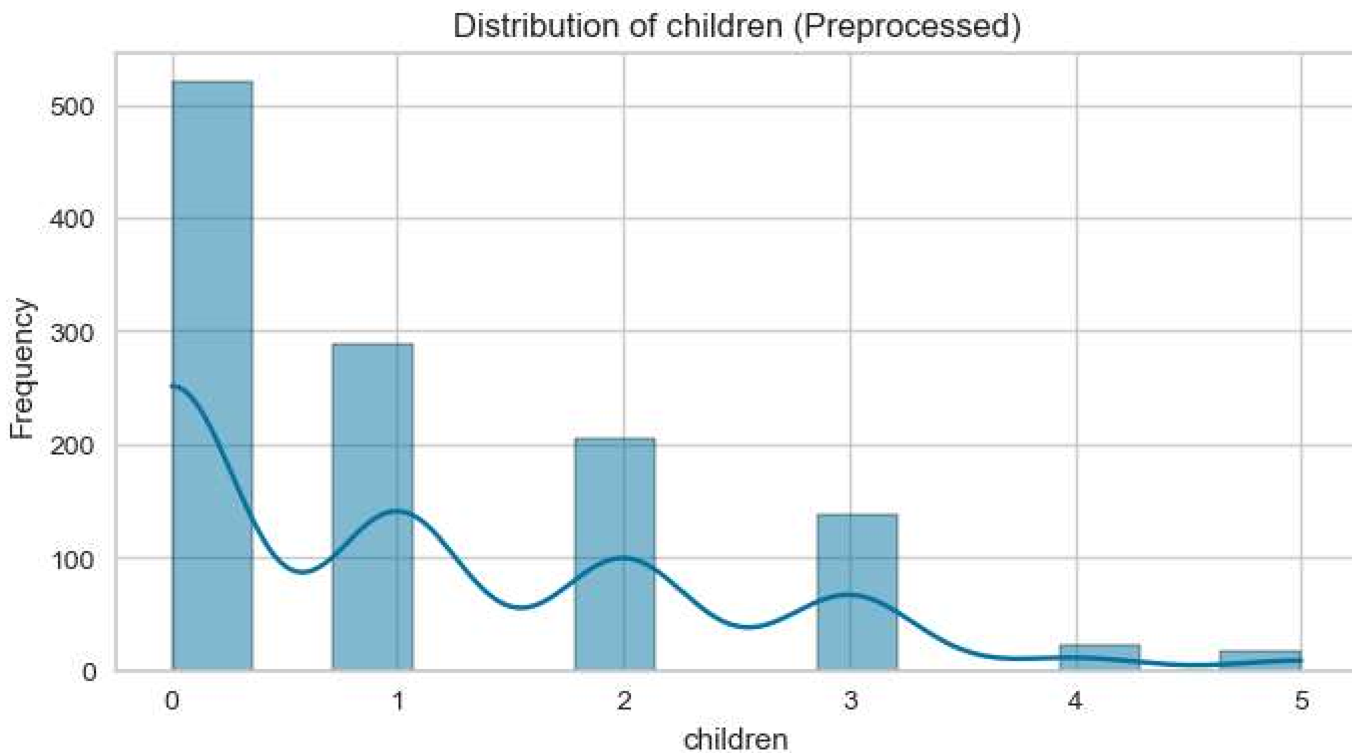

Distribution of sex (Original)
700
600
500
400
300
200
100
0
Count
female
male
sex

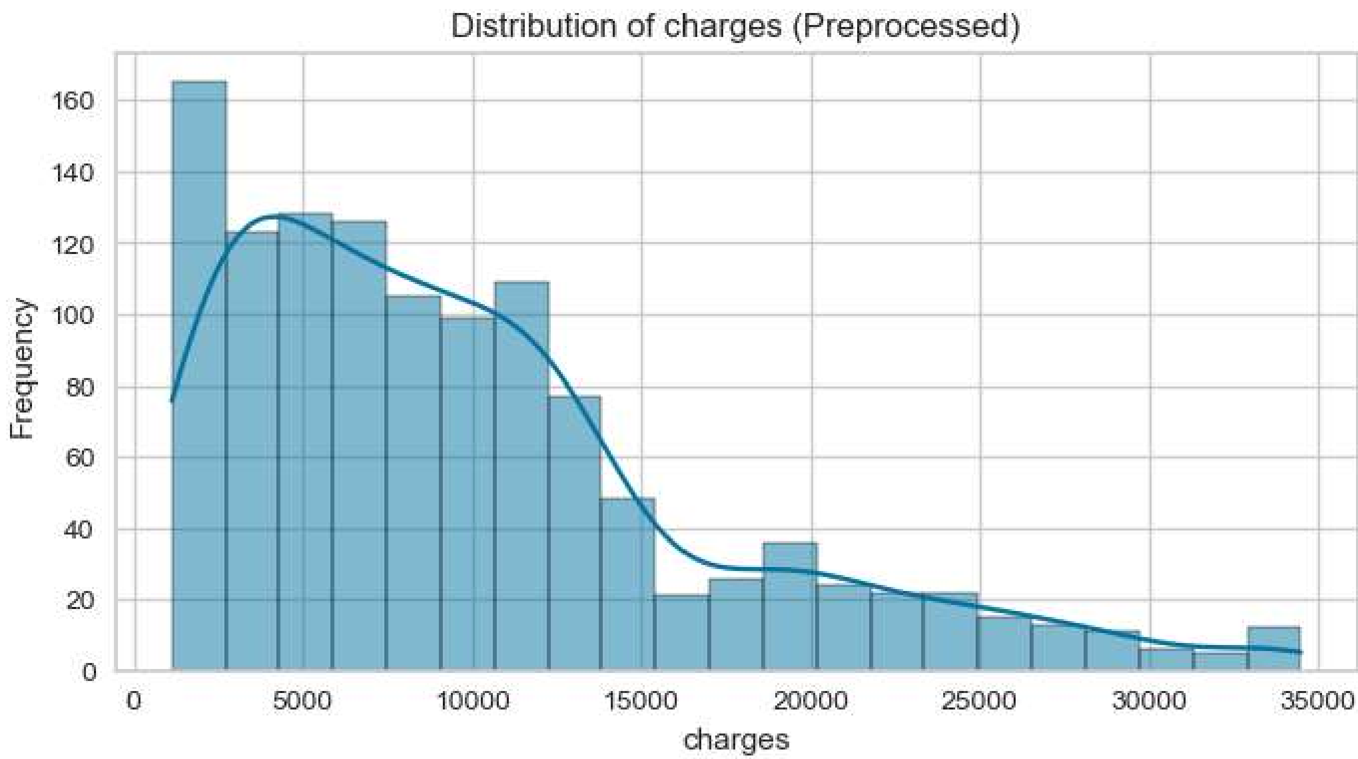

Distribution of smoker (Original)
Count
1000
800
600
400
200
0
yes
no
smoker

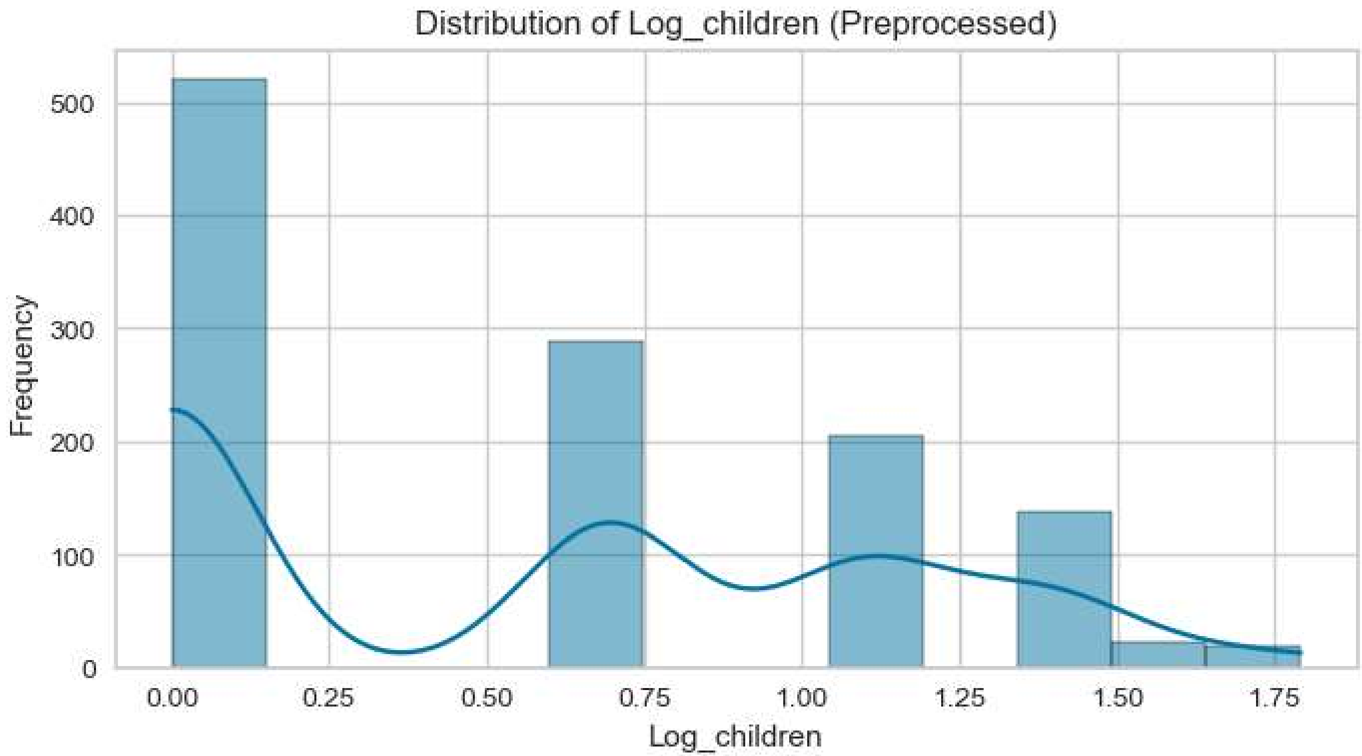

Distribution of region (Original)
Count
350
300
250
200
150
100
50
0
southwest
southeast
northwest
northeast
region


Distribution of age (Preprocessed)
Frequency
175
150
125
100
75
50
25
0
20
30
40
50
60
age

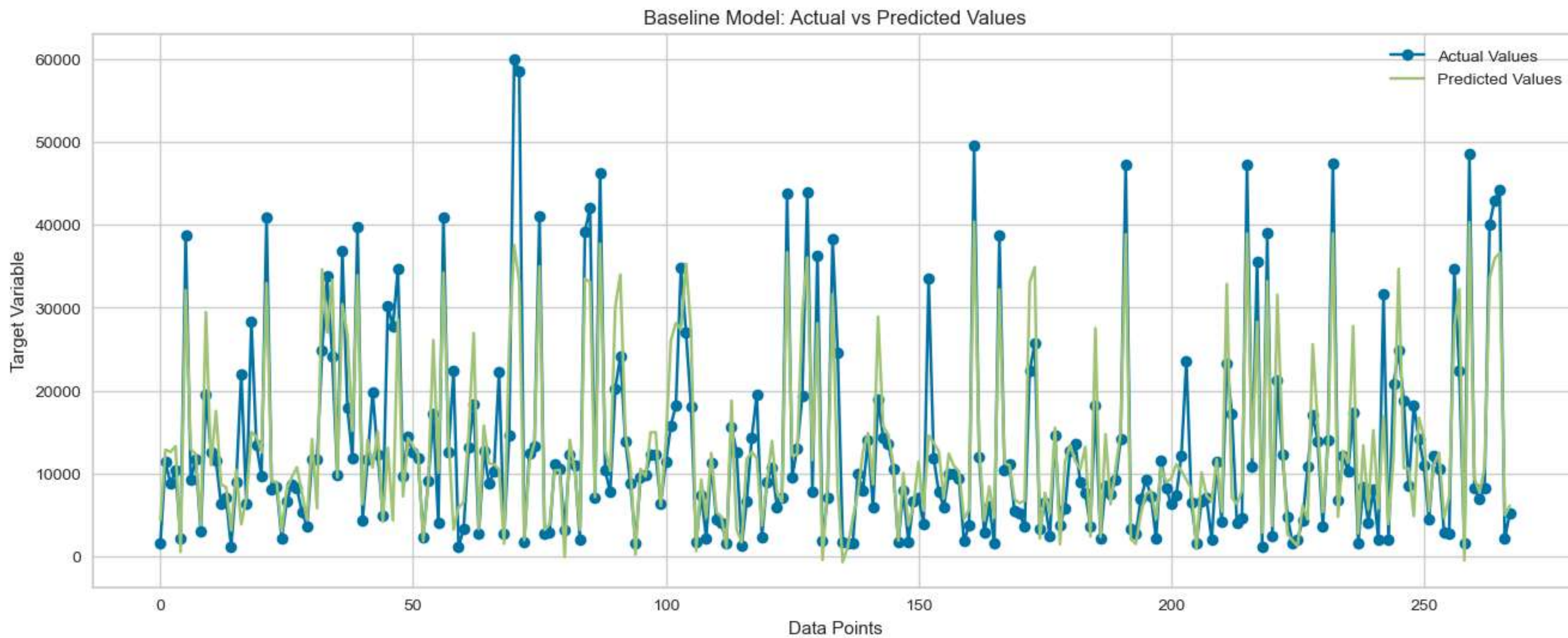

Distribution of bmi (Preprocessed)
Frequency
0
20
40
60
80
100
120
15
20
25
30
35
40
45
bmi

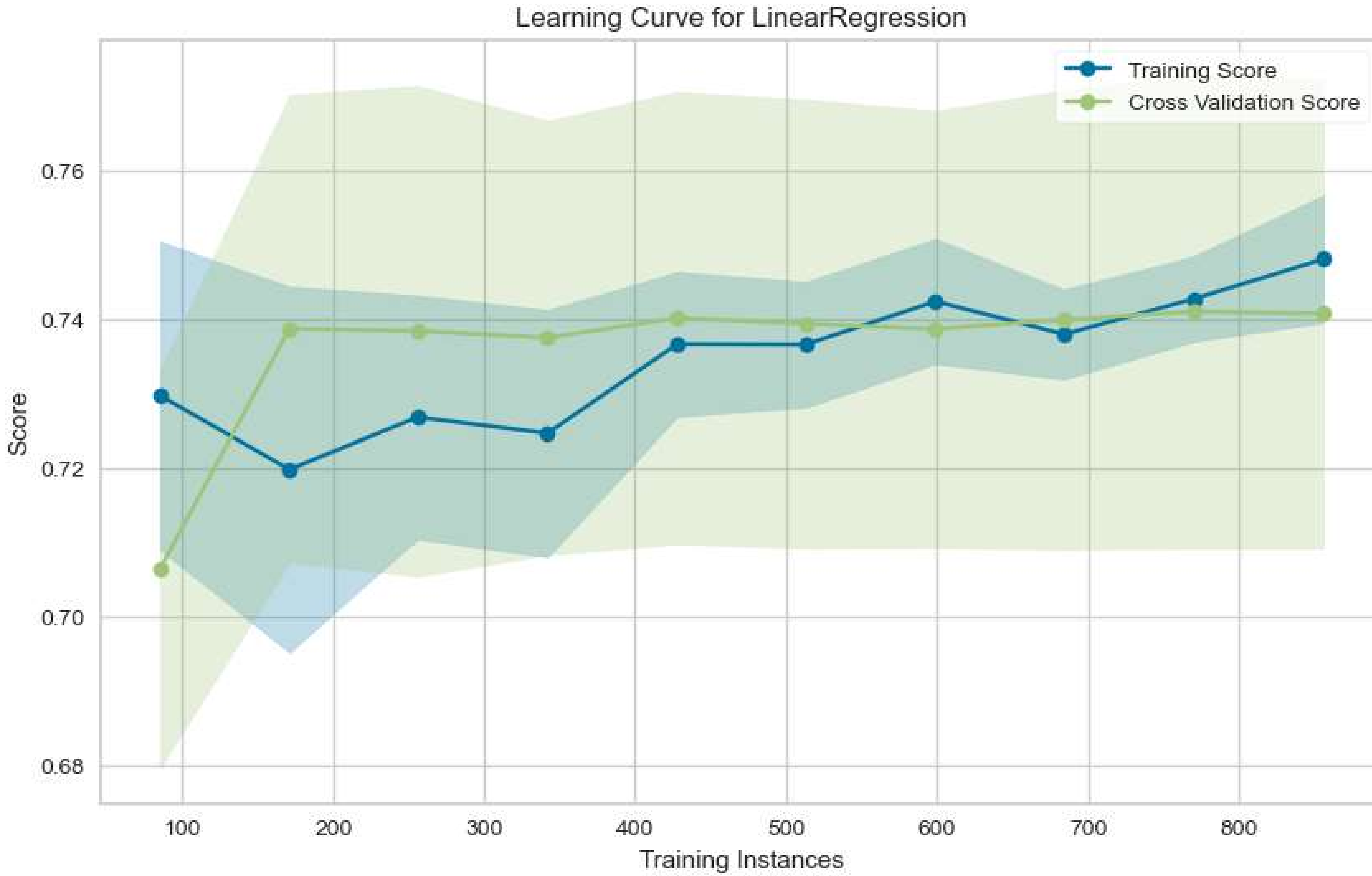

Distribution of children (Preprocessed)
Frequency
0
100
200
300
400
500
0
1
2
3
4
5
children


Distribution of charges (Preprocessed)
Frequency
0
20
40
60
80
100
120
140
160
0
5000
10000
15000
20000
25000
30000
35000
charges

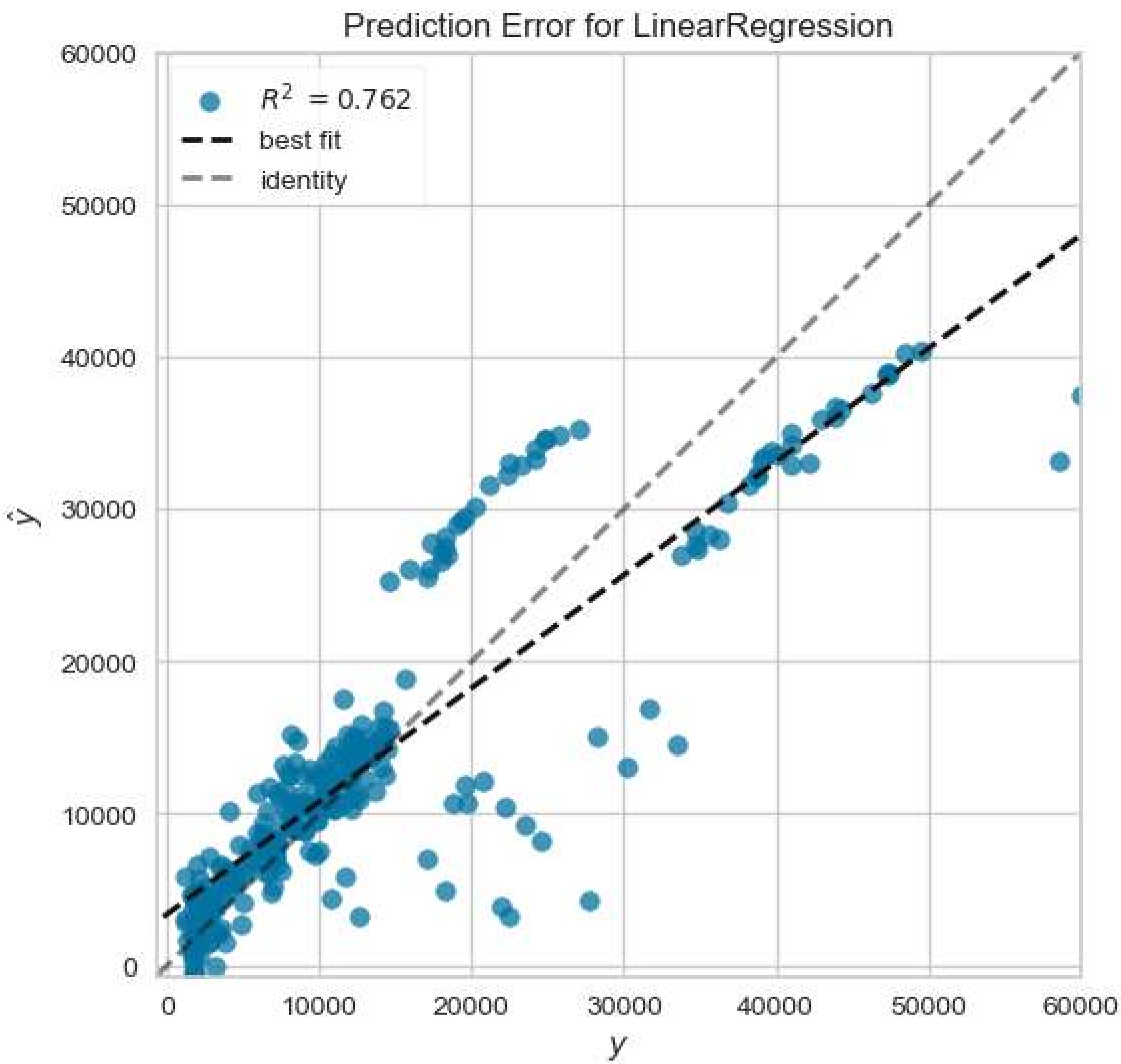


[Step 2] Training and Evaluating Baseline Model

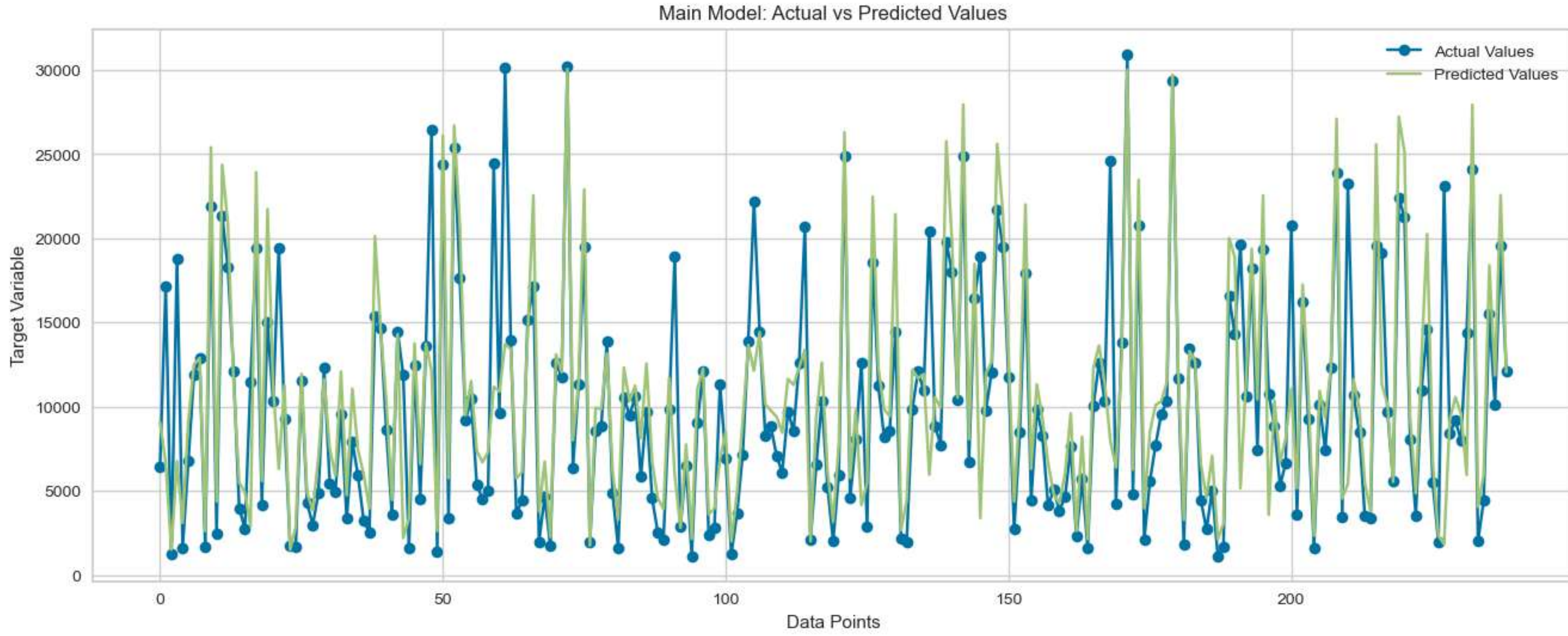

Baseline Model: Actual vs Predicted Values
Actual Values
Predicted Values
Target Variable
Data Points
0
10000
20000
30000
40000
50000
60000
0
50
100
150
200
250

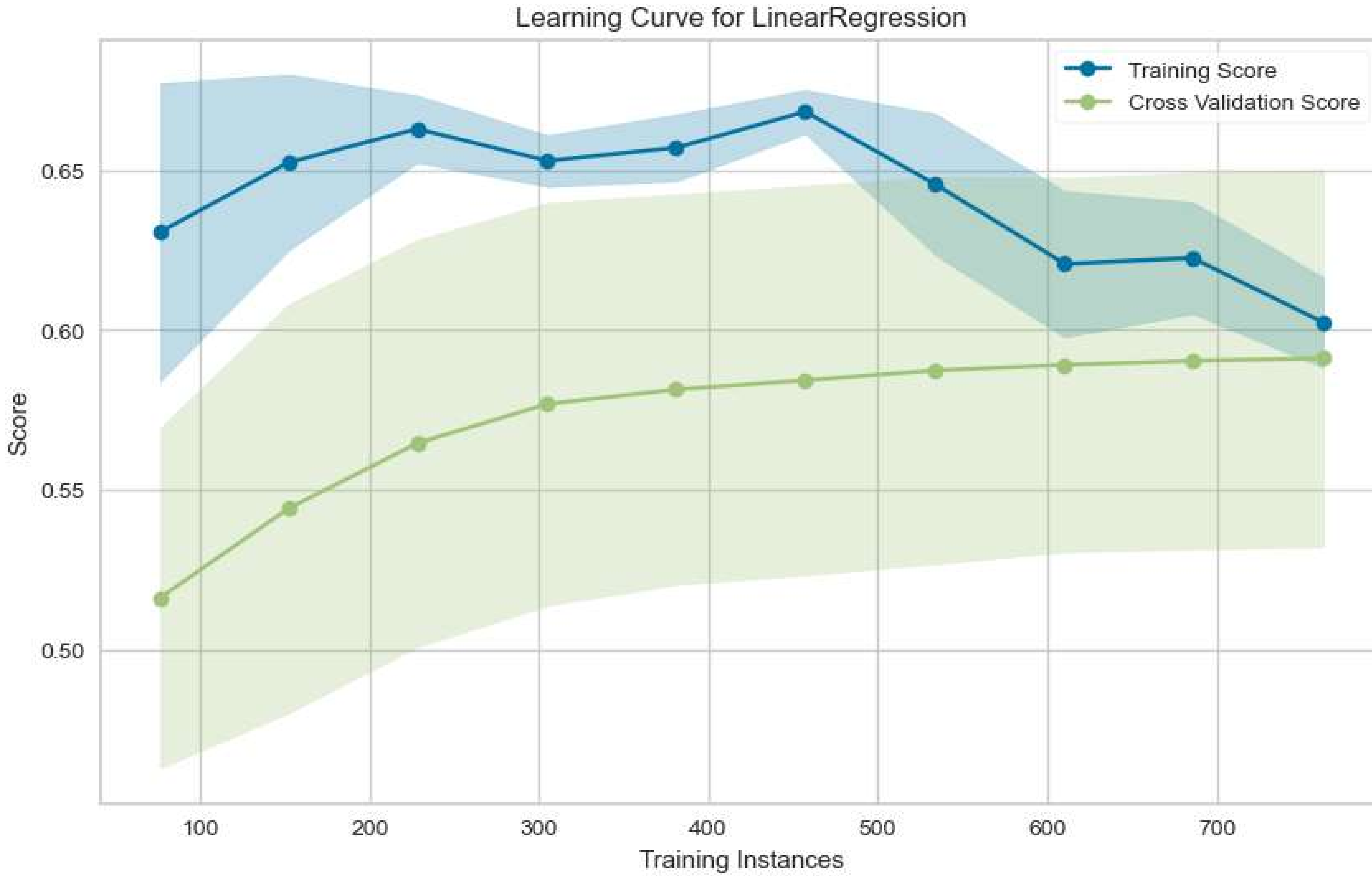



```
C:\Users\vaibh\anaconda3\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but L
inearRegression was fitted with feature names
  warnings.warn(
```

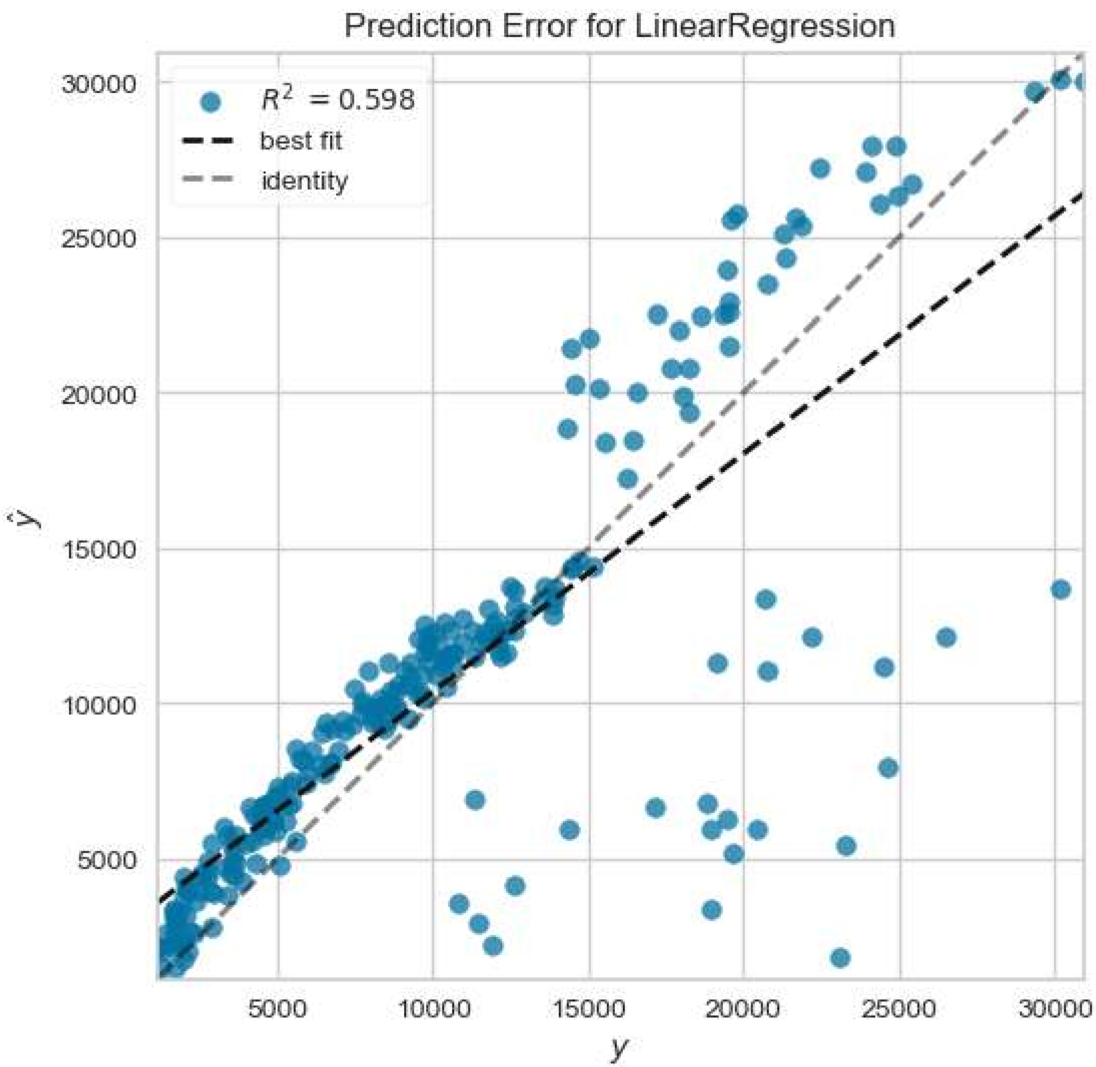


[Step 3] Training and Evaluating Main Model

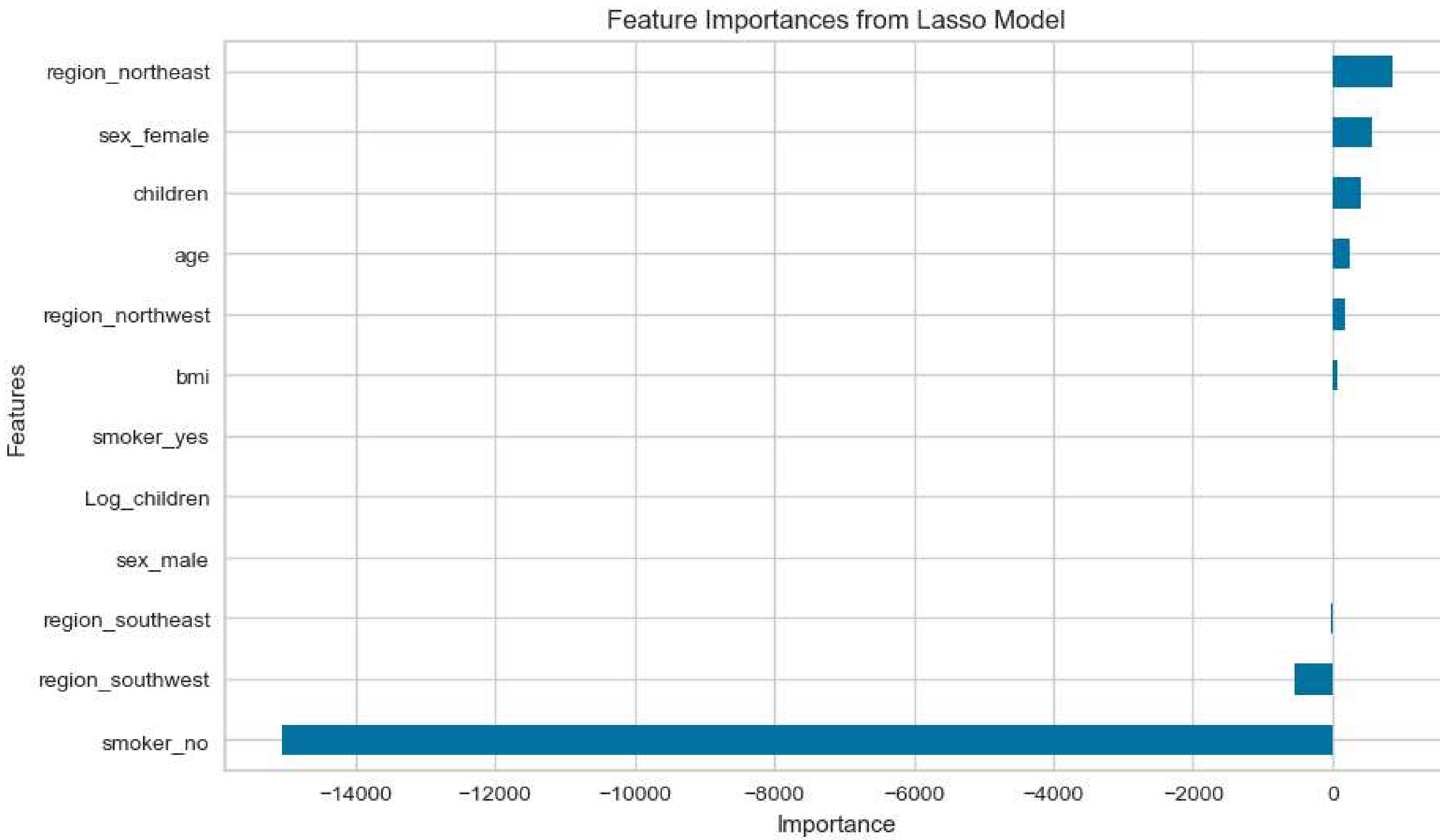

Main Model: Actual vs Predicted Values
Actual Values
Predicted Values
Target Variable
Data Points
0
5000
10000
15000
20000
25000
30000
0
50
100
150
200




```
C:\Users\vaibh\anaconda3\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but L
inearRegression was fitted with feature names
  warnings.warn(
```



[Step 4] Performing Lasso Feature Importance Analysis




```
Best alpha: 43.87988569558819
MSE: 18508109.839097515
R-squared: 0.61

===== Model Comparison Results =====
Baseline Model:
  - MSE: 35479352.807
  - R-squared: 0.762
  - MAE: 2673.507
Main Model:
  - MSE: 19229100.882
  - R-squared: 0.598
  - MAE: 4051.859
====================================
```

In [ ]:

```

```