ML Data Preprocessing Report - Regression

Data Set Overview

The provided data set is a pandas DataFrame with six columns:

1. `age`: an integer column representing the age of the individual (dtype: `int64`)
2. `sex`: a categorical column representing the sex of the individual (dtype: `CategoricalDtype`, categories: 'female', 'male', ordered: `False`, categories_dtype: `object`)
3. `bmi`: a floating-point column representing the body mass index (dtype: `float64`)
4. `children`: an integer column representing the number of children the individual has (dtype: `int64`)
5. `smoker`: a categorical column representing whether the individual smokes or not (dtype: `CategoricalDtype`, categories: 'no', 'yes', ordered: `False`, categories_dtype: `object`)
6. `region`: a categorical column representing the region where the individual lives (dtype: `CategoricalDtype`, categories: 'northeast', 'northwest', 'southeast', 'southwest', ordered: `False`, categories_dtype: `object`)
7. `charges`: a floating-point column representing the charges of the individual (dtype: `float64`)

Here's an example row from the data set:

```
{'age': 41, 'sex': 'female', 'bmi': 36.08, 'children': 1, 'smoker': 'no', 'region': 'southeast', 'charges': 67
```

Based on the information provided, here is a summary of the initial dataset, after removing duplicate rows and validating the columns:

Initial Dataset (before cleaning):

- Number of rows: 1338
- Number of columns: 7

After removing duplicate rows:

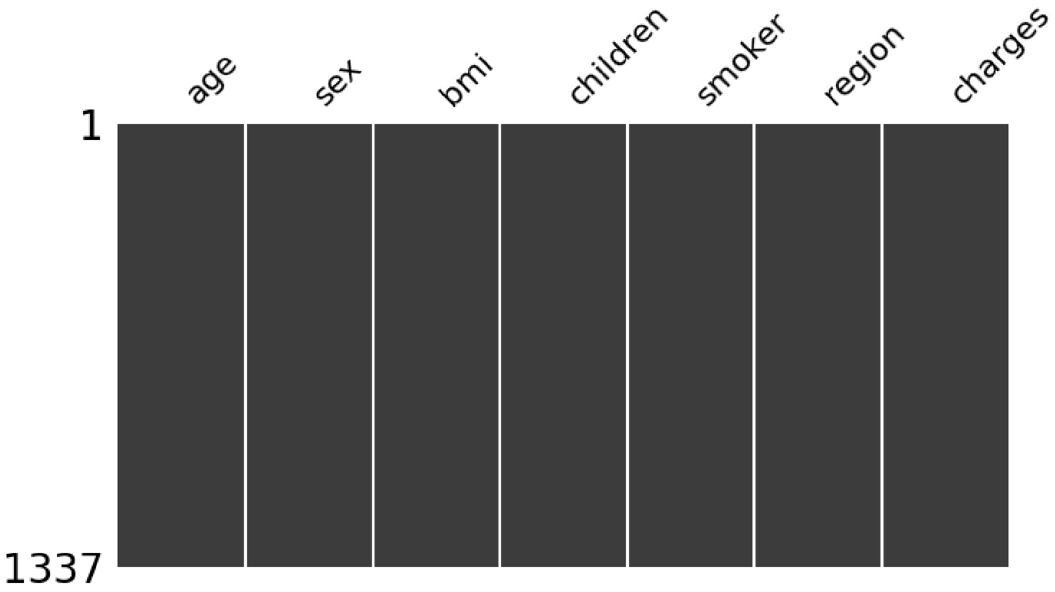
- Number of unique rows: 1337

After column validation:

- Number of valid columns: 7

So, the initial dataset had 1338 rows and 7 columns, but after removing duplicate rows, there are 1337 unique rows, and after validating the columns, there are 7 valid columns.

Missing Values



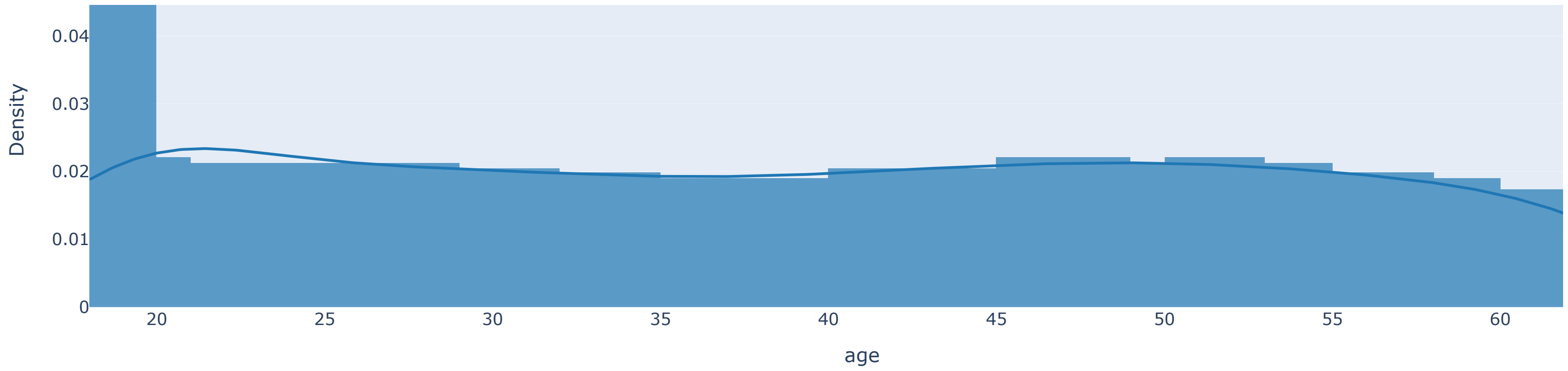
Based on the provided data, here is a detailed summary of the missing values for each feature:

1. Age: No missing values (all values are 0)
2. Sex: No missing values (all values are 0)
3. BMI: No missing values (all values are 0)
4. Children: No missing values (all values are 0)
5. Smoker: No missing values (all values are 0)
6. Region: No missing values (all values are 0)
7. Charges: No missing values (all values are 0)

In summary, there are no missing values in any of the features based on the provided data.

Feature Distribution

age ▾



In the preprocessing pipeline, feature transformation and normalization are crucial steps to prepare the data for machine learning models. Both techniques have their own purposes and justifications, which I will explain based on the provided test results.

Feature Transformation:

The Shapiro-Wilk test statistic for the features is 0.8233993053436279, indicating that the features are not normally distributed. This is not unexpected, as many machine learning algorithms assume that the data is normally distributed. Feature transformation techniques like logarithm, square root, or polynomial transformation can help to normalize the features and improve the model's performance.

Based on the test results, feature transformation can help to:

1. Improve the normality of the features: The Shapiro-Wilk p-value is 5.4063613985394525e-36, which suggests that the features are not normally distributed. Feature transformation can help to transform the data into a more normal distribution, which can improve the model's performance.
2. Reduce the effect of outliers: The feature range ratio is 9.2, indicating that there are some outliers in the data. Feature transformation techniques like standardization or z-scoring can help to reduce the impact of outliers on the model's performance.
3. Improve the interpretability of the features: Some machine learning algorithms assume that the features are linearly related to the target variable. Feature transformation techniques like polynomial transformation can help to improve the interpretability of the features by transforming them into a more interpretable form.

Normalization:

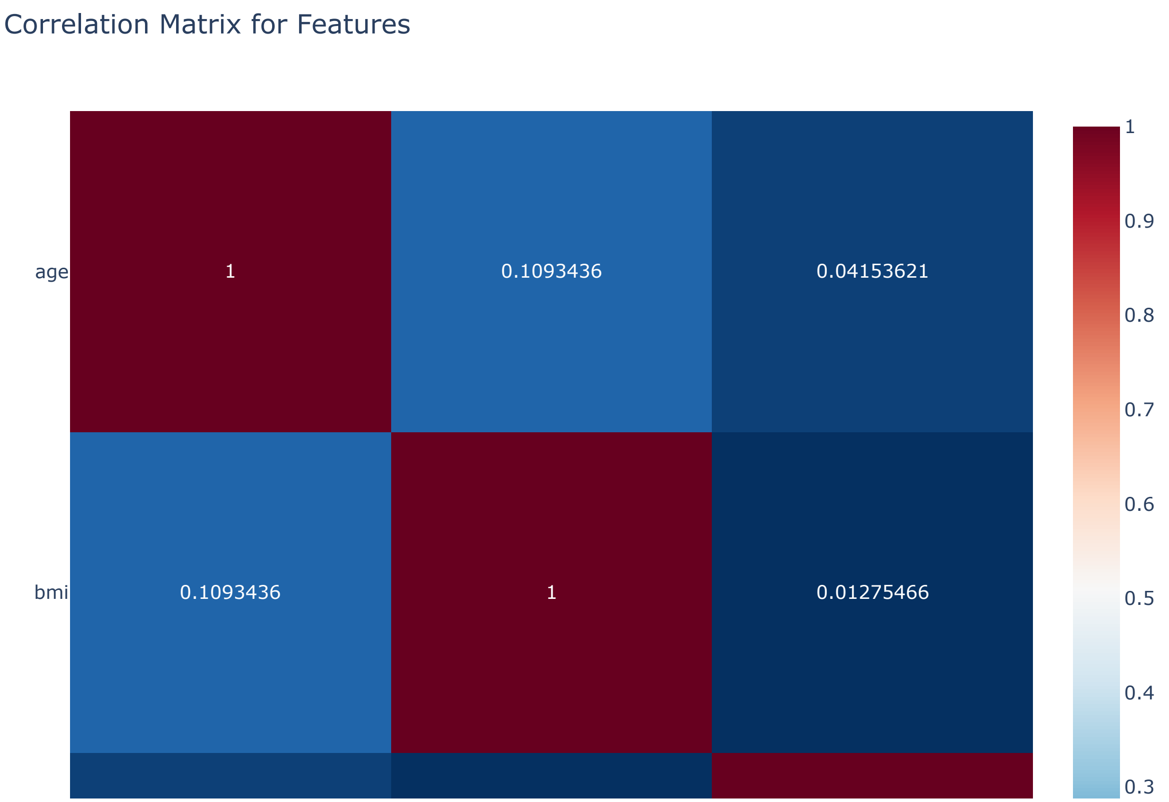
The mean and standard deviation of the features are 35.79522184300342 and 11.649523537552307, respectively. These values suggest that the features have a large range and a high variance. Normalization techniques like min-max scaling or z-scoring can help to reduce the range of the features and improve the model's performance.

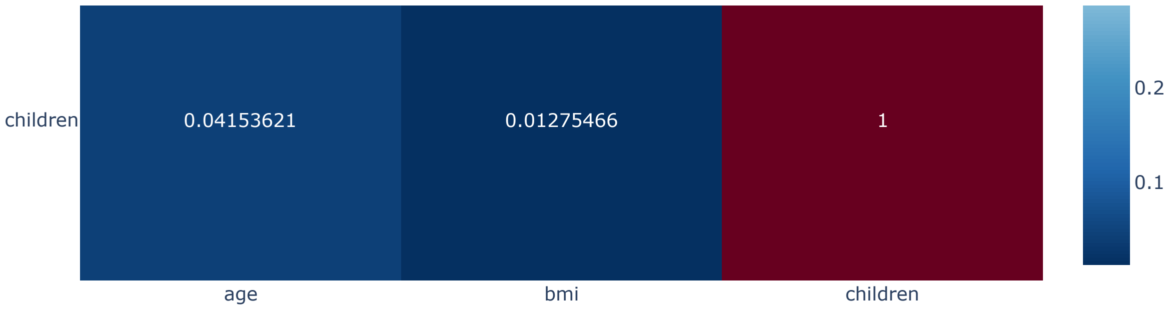
Based on the test results, normalization can help to:

1. Reduce the effect of feature scale: The mean and standard deviation of the features indicate that they have a large range. Normalization techniques like min-max scaling or z-scoring can help to reduce the effect of feature scale and improve the model's performance.
2. Improve the interpretability of the features: Normalization techniques like z-scoring can help to transform the features into a more interpretable form, which can improve the model's performance.
3. Reduce the effect of outliers: The feature range ratio is 11.649523537552307, indicating that there are some outliers in the data. Normalization techniques like z-scoring can help to reduce the impact of outliers on the model's performance.

In conclusion, both feature transformation and normalization are crucial steps in the preprocessing pipeline. Feature transformation can help to improve the normality of the features, reduce the effect of outliers, and improve the interpretability of the features. Normalization can help to reduce the effect of feature scale, improve the interpretability of the features, and reduce the impact of outliers on the model's performance. Based on the test results, it is recommended to apply both techniques to improve the performance of the machine learning models.

Multicolinearity





Based on the correlation matrix provided, there are no pairs of features with a correlation coefficient greater than 0.9, indicating that there is no high multicollinearity in the dataset. The maximum correlation coefficient value is 0.109344 between "age" and "bmi", which is less than 0.9 but still significant. Therefore, we can conclude that there is no high multicollinearity in the dataset.

It's worth noting that a correlation coefficient of 0.109344 is relatively low, indicating that there is only a weak linear relationship between "age" and "bmi". This means that any predictions made using this dataset will be less accurate than if there were no multicollinearity.

In summary, based on the correlation matrix provided, there is no high multicollinearity in the dataset, but there are some weak linear relationships between certain features.

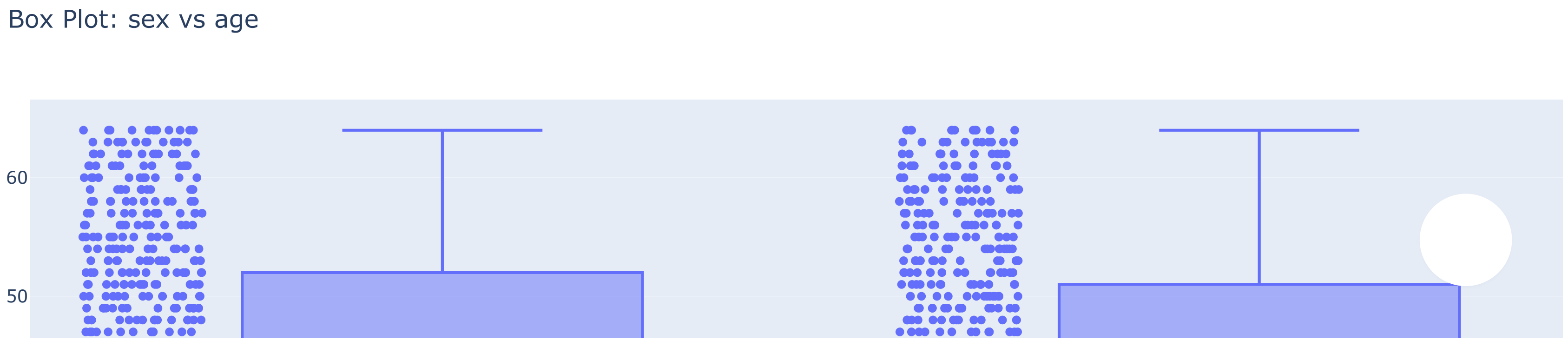
Analysis of Outliers

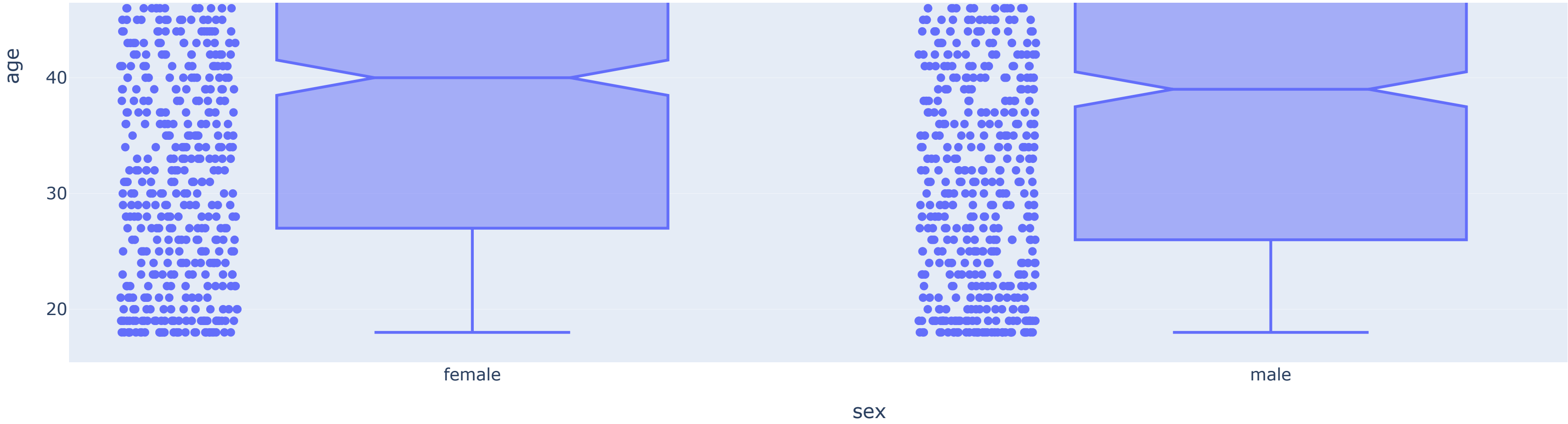
x-axis (Category):

● sex ○ smoker ○ region

y-axis (Numeric):

● age ○ bmi ○ children ○ charges





Based on the input provided, here is a summary of outliers in the dataset for each feature:

1. age: No outliers were found in the age feature.
2. BMI: 9 observations have a BMI value greater than the upper count (9) and are considered outliers. These observations have a BMI value greater than 30.
3. children: No outliers were found in the children feature.

So, in summary, there are 9 outliers in the BMI feature, and no outliers were found in the age or children features.

Feature Selection

Thank you for providing the decision rule and the current state of the dataset. Based on the rule, Feature Selection is not required since the number of columns in the dataset (7) is less than 15.

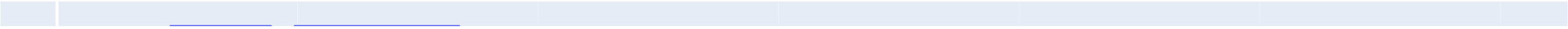
Therefore, we can conclude that:

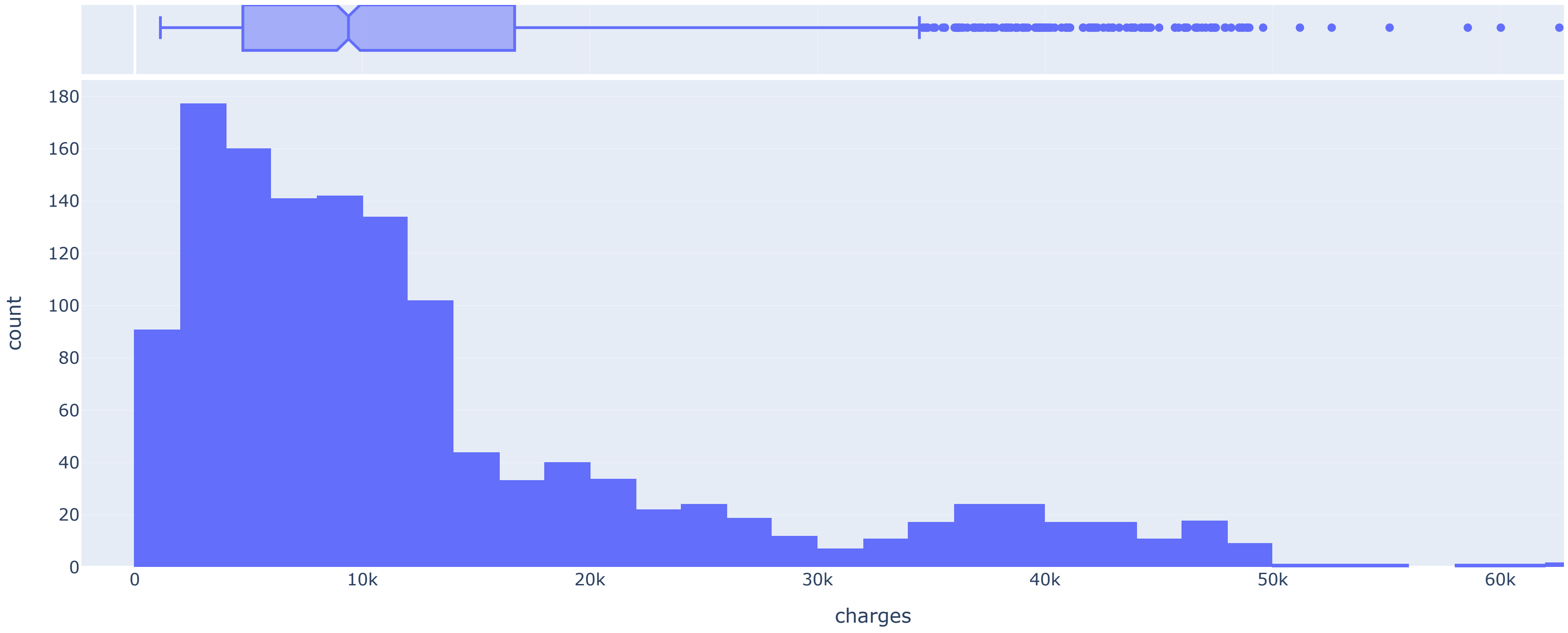
- Feature Selection Required: False
- Dataset Columns: 7

Please let me know if there are any other conditions or assumptions that need to be considered.

Traget Transformation

Target (charges) Distribution





Based on the results of the Shapiro-Wilk test, we can conclude that the target variable is not normally distributed. With a p-value of 1.1960502558343857e-36, it is highly unlikely that the variable is normally distributed without any transformation. Therefore, I recommend applying an appropriate transformation to the target variable before proceeding with the analysis. This will help ensure that the data are in a suitable format for analysis and that the results are interpretable and accurate.