**BDNS ASSIGNMENT:**

**TOPIC:**

Predicting Medical Insurance Charges Using Mongo dB and Spark.

**NAME: SUMIT DUTTA**

**ROLL NO: C23034**

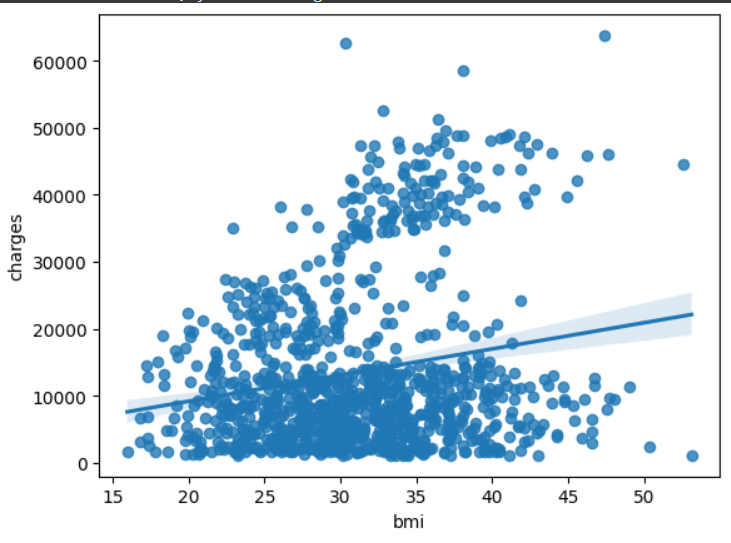
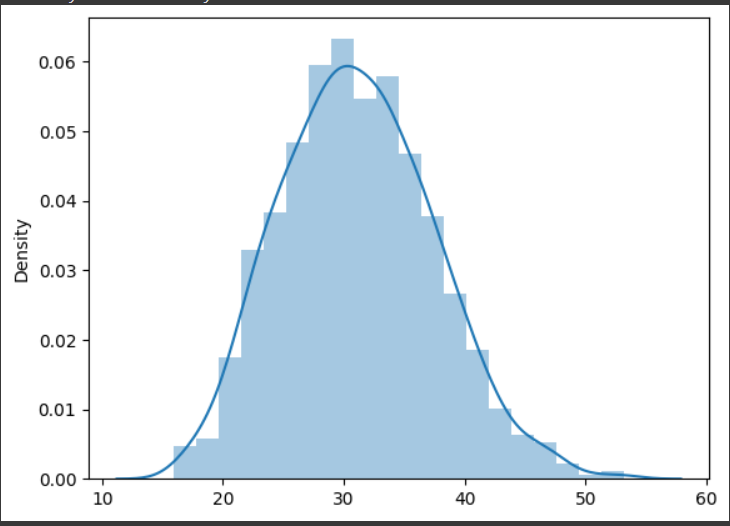
**OBJECTIVE:**

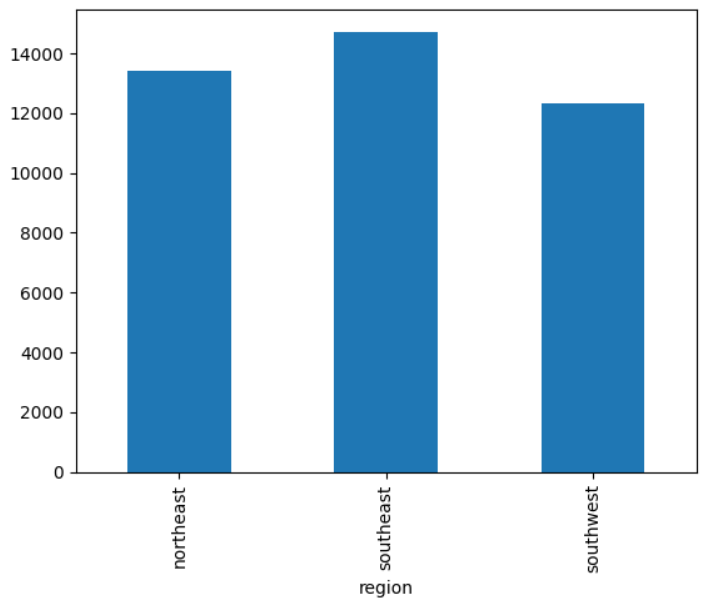
In this project focused on predicting medical insurance charges, our aim was to build a model using the Spark distributed data processing framework. The model would estimate the charges individuals would incur based on factors such as age, gender, BMI (Body Mass Index), smoking status, region, and number of children. To efficiently manage and analyze the medical insurance data, we integrated MongoDB with Spark.

**TARGET: Medical Insurance charge.**

**PROCEDURE:**

During the exploratory data analysis (EDA) phase, we utilized various data visualization charts and plots to analyze the data both individually and in combination. Our objective was to gain insights into variables such as transmission type, fuel type, car name, owner, and manufacturing year. By visually examining these variables, we aimed to identify patterns, outliers, and relationships within the data. These insights played a crucial role in guiding the subsequent steps of the project

****



* Based on the bivariate analysis conducted for medical insurance charge prediction, it was observed that there is a negative correlation between age and the predicted insurance charges. This indicates that as age increases, the estimated insurance charges tend to decrease.
* Moreover, the analysis of different factors such as BMI (Body Mass Index), smoking status, and number of children revealed interesting insights. For example, individuals with a higher BMI tend to have higher predicted insurance charges. Similarly, smokers are associated with higher estimated charges compared to non-smokers. Additionally, having a greater number of children is positively correlated with higher insurance charges.
* It is important to note that these observations are based on the analysis of the available data and may not necessarily imply causation.
* There is a negative correlation of -0.043 between children and charges. This suggests that as the number of children dependent on the individual increases, the insurance charges tends to decrease.
* There is a positive correlation of 0.19 between BMI and the charges. This indicates that as the BMI of the individual increases, the insurance charges tends to increase.

**DATA PREPARATION:**

* Conversion of categorical columns to numerical: The categorical columns were transformed into numerical columns using the One-Hot Encoder technique. This conversion enables the representation of categorical data in a numerical format suitable for analysis.
* Usage of String Indexer: The String Indexer class was employed to encode categorical variables, such as the “region" column. This process assigns unique numerical indices to each distinct category within the column. The resulting indexed values are stored in a new column called “region\_type\_indexer.“
* Removal of unnecessary columns: To streamline the dataset, the redundant categorical columns were dropped since they were replaced with indexed and vector columns. This helps in eliminating duplicate information and improving the efficiency of subsequent analysis.

PIPELINE CREATION:

The creation of pipeline stages involves setting up a pipeline comprising two stages: a type indexer and a type encoder. These stages use transformers, namely the Type\_Indexer and Type\_Encoder, to preprocess the dataset. Once the pipeline is defined, the fit() method is applied to the pipeline object using the new\_data dataset as input. This trains the pipeline and produces a fitted pipeline (pipeline\_model) that can be utilized to transform new data.

In addition, a Standard Scaler is employed to scale the features within a consistent range. The Standard Scaler ensures that each value is scaled to a range between 0 and 1, enabling fair comparisons and reducing the impact of varying feature magnitudes.

**TRAINING THE MODEL:**

* The train-test split involves dividing the scaleddf dataset into two separate datasets: the training dataset and the test dataset.
* This split is achieved using the randomSplit() method, which takes two parameters: weights and seed.
* The weights parameter determines the relative sizes of the resulting datasets, while the seed parameter is optional and used for reproducibility purposes.
* In this case, the training dataset is allocated 70% of the data, while the test dataset receives 30% of the data.

**OUTPUT:**

The training data is used to train the linear regression model, and then the test data is transformed using the same model to make predictions. The model coefficients and intercept are calculated to interpret the relationships between predictors and the target variable, enabling predictions and performance evaluation.

However, the R-squared value obtained from the model, which is 0.63, suggests that the model's fit is not satisfactory. This indicates that only 63% of the variability in the target variable is explained by the predictors included in the model. To thoroughly evaluate the model and identify potential areas for improvement, further investigation is necessary.