# Health Insurance Cost Prediction

- Maganti Manogna Venkata Sudha Anirudh

# Abstract

The escalating costs, limited accessibility, and scope of medical insurance underscore the need for accurate cost prediction models. Addressing this imperative can revolutionize financial planning for individuals and businesses, empower fair premium setting for insurers, and refine risk management strategies. However, conventional frequentist approaches exhibit limitations in flexibility, interpretability, and predictive capacity, necessitating exploration of alternative methodologies.

This study delves into the Bayesian paradigm as an alternative approach to model medical insurance costs. Leveraging Bayesian modeling techniques, including Markov Chain Monte Carlo (MCMC) and Laplace Approximation, alongside computational methodologies, this research aims to construct robust predictive models. By employing various regression models on comprehensive medical insurance datasets, the study seeks to unravel the intricate relationship between diverse parameters and insurance costs.

The Bayesian framework offers distinct advantages over conventional methods, embracing uncertainty, accommodating prior knowledge, and enhancing interpretability. By sidestepping the restrictive assumptions of frequentist models, this approach promises a more nuanced understanding of cost determinants and superior predictive capabilities.

The incorporation of alternative methodologies, particularly Bayesian modeling, into the realm of medical insurance cost prediction represents a significant step toward addressing the complexities and uncertainties inherent in insurance economics. This research endeavors to contribute valuable insights that can reshape the landscape of insurance cost prediction, fostering more informed decision-making and promoting financial stability in the healthcare domain.

# Introduction

In the complex landscape of insurance pricing, the quest for precision and reliability in predicting costs has become increasingly vital. Yet, the conventional methods rooted in frequentist approaches reveal inherent limitations, relying heavily on rigid data assumptions and lacking the adaptability required in today's dynamic insurance ecosystem. Recognizing these constraints, there's a burgeoning interest in alternative modeling methodologies, particularly in the realms of Bayesian modeling and machine learning.

The significance of accurately predicting insurance costs cannot be overstated. It serves as a linchpin for individuals and businesses alike, enabling prudent financial planning and risk mitigation strategies. Moreover, for insurance companies, the ability to set fair premiums while remaining competitive is contingent upon robust predictive models.

This project embarks on a journey leveraging Bayesian methodology, specifically employing Markov Chain Monte Carlo (MCMC) techniques and computational tools, to model insurance prices. While recognizing the importance of traditional regression models like Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor, this study seeks to expand the horizon by delving deeper into the Bayesian approach.

The focus here extends beyond conventional approaches by embracing Bayesian Linear Regression and Bayesian Multiple Linear Regression. The intent is to compare and contrast these Bayesian models against the aforementioned regression models, gauging their efficacy based on metrics like R Square and RMSE. By doing so, we aim to uncover the strengths and advantages of Bayesian modeling in predicting insurance costs, potentially offering a more nuanced and accurate understanding that transcends the limitations of traditional frequentist techniques.

# Problem Description

Medical insurance is vital for financial protection, but its cost, limited availability, and coverage pose significant challenges. To overcome these, developing a model for predicting insurance costs accurately can be helpful.

Such a model can help individuals and businesses budget and plan accordingly, enable insurance companies to set fair premiums and remain competitive, and improve risk management strategies. However, frequentist approaches have limitations, such as reliance on assumptions about data distribution and lack of flexibility, interpretability, and predictive power.

Alternative modeling approaches like Bayesian modeling or machine learning may offer additional benefits. Developing a model for predicting insurance costs is crucial in today's world, and alternative approaches can be considered to overcome the limitations of frequentist approaches.

In this project, we look at the Bayesian approach for modeling insurance prices using Markov Chain Monte Carlo (MCMC), Laplace Approximation and some computational techniques and fitting various Regression models with medical insurance data.

# Data Description

The dataset comprises of several crucial features. It includes the age of insured individuals, their gender ('female' or 'male'), their Body Mass Index (BMI), which serves as a metric for assessing body weight, the number of children or dependents they have, their smoking status (categorized as 'yes' for smokers and 'no' for non-smokers), the geographic region to which they belong (with categories like 'southwest,' 'southeast,' 'northeast,' and 'northwest'). The final column, "charges," quantifies the health insurance costs for each policyholder, serving as the target variable for predictive modeling. The dataset comprises 1338 entries with 7 essential features.

```
      age     sex     bmi  children smoker     region      charges
0      19  female  27.900         0    yes  southwest  16884.92400
1      18    male  33.770         1     no  southeast   1725.55230
2      28    male  33.000         3     no  southeast   4449.46200
3      33    male  22.705         0     no  northwest  21984.47061
4      32    male  28.880         0     no  northwest   3866.85520
...   ...     ...     ...       ...    ...        ...         ...
1333   50    male  30.970         3     no  northwest  10600.54830
1334   18  female  31.920         0     no  northeast   2205.98080
1335   18  female  36.850         0     no  southeast   1629.83350
1336   21  female  25.800         0     no  southwest   2007.94500
1337   61  female  29.070         0    yes  northwest  29141.36030

[1338 rows x 7 columns]
```

# Exploratory Data Analysis (EDA)

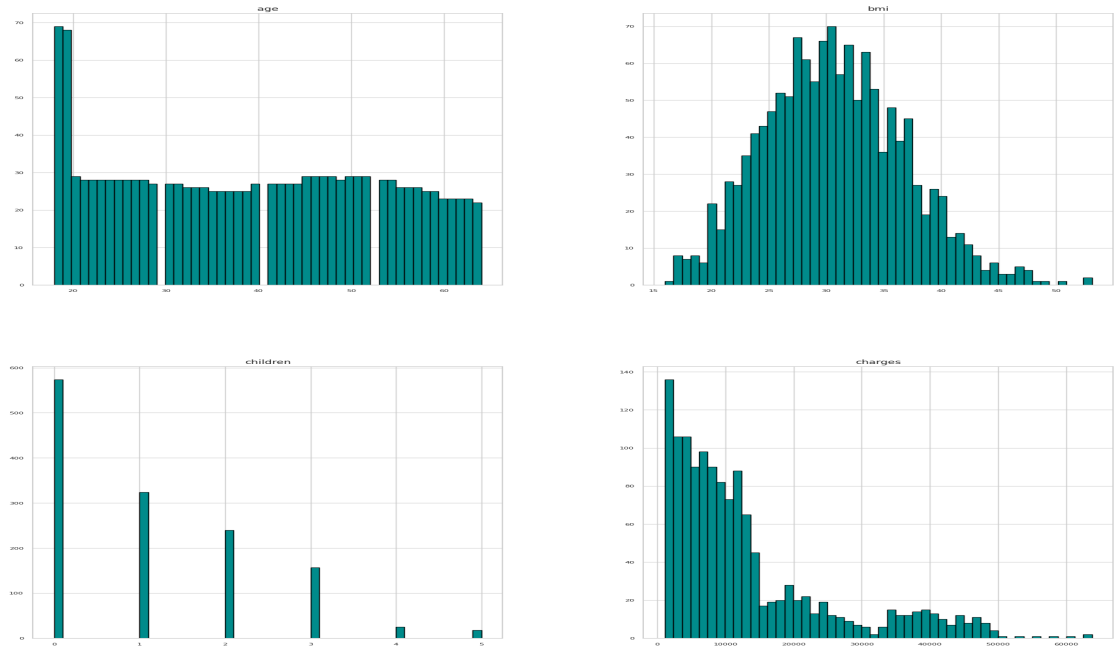
Exploratory Data Analysis Overview:
In the initial phase of exploration, a dataset of dimensions 1338x7 was loaded, revealing no missing values across the entirety of the dataset.

1. Unveiling Numerical Insights: You initiated the EDA by delving into the numerical features, dissecting their distributions and characteristics. By scrutinizing statistical attributes like mean, median, and standard deviation, you gained insights into the central tendencies and spread of the data. Understanding the nature of these distributions, whether they adhere to a normal distribution or exhibit skewness, provided vital context for subsequent analyses and model assumptions.

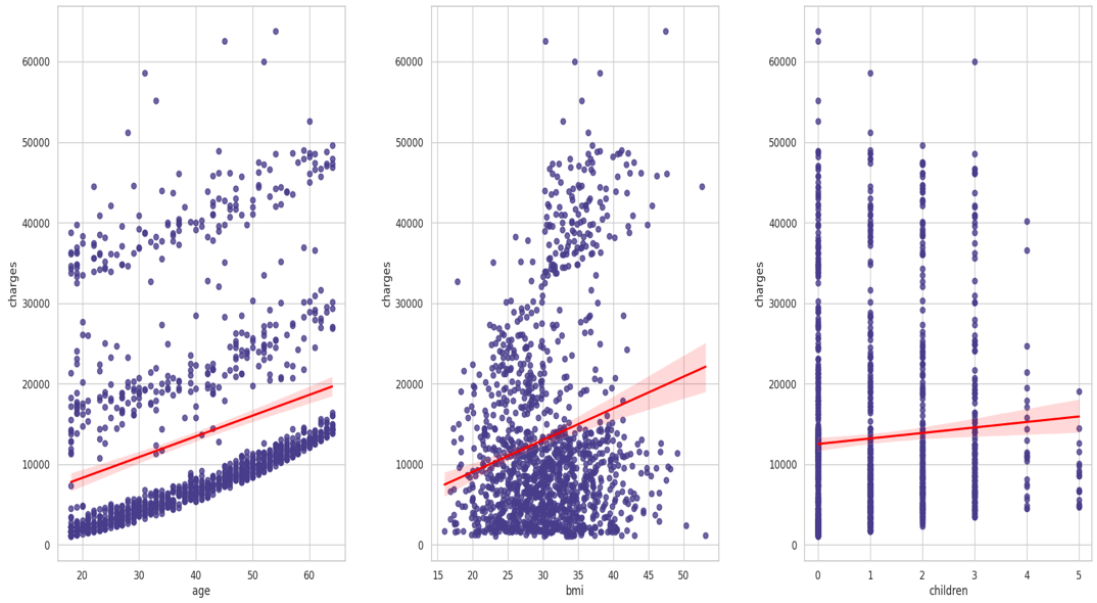
| | age | bmi | children | charges |
|---|---|---|---|---|
| **0** | 19 | 27.900 | 0 | 16884.92400 |
| **1** | 18 | 33.770 | 1 | 1725.55230 |
| **2** | 28 | 33.000 | 3 | 4449.46200 |
| **3** | 33 | 22.705 | 0 | 21984.47061 |
| **4** | 32 | 28.880 | 0 | 3866.85520 |

2. Unravelling Relationships: The EDA continued with an exploration of relationships between key numerical attributes and the target variable, insurance charges. Through scatterplots, you visually depicted the interplay between age, BMI, and the number of children against insurance costs. This step enabled a nuanced understanding of how these factors potentially influence insurance charges, laying the groundwork for informed modeling strategies.

Feature Distribution:

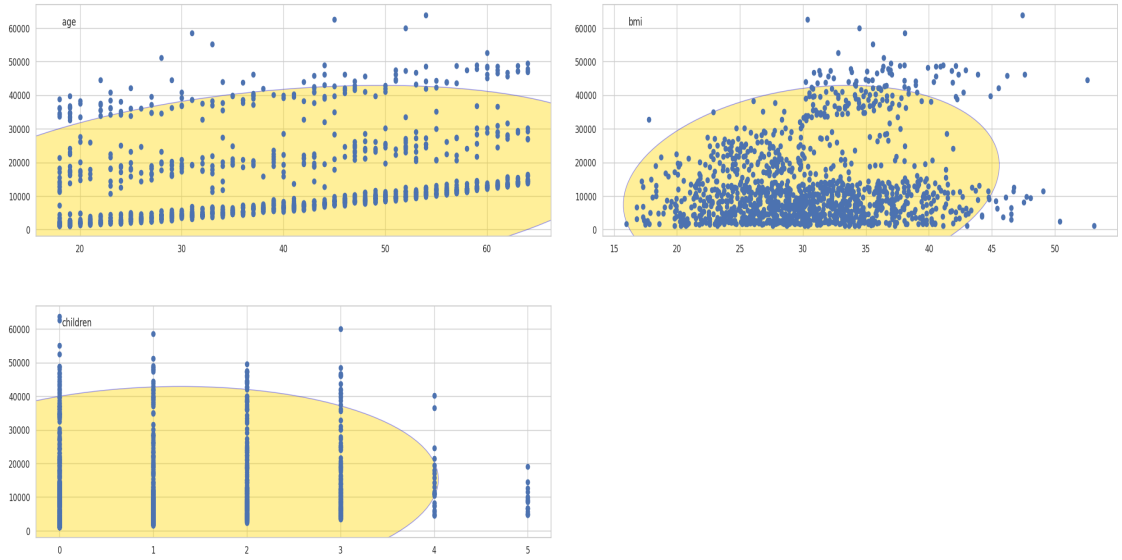



Scatterplot & Relationship Assessment:




3. Bivariate Outlier Detection: Employing Mahalanobis Distance for bivariate outlier detection showcased a meticulous approach to identify anomalies that deviate significantly from the expected distribution patterns. This method, leveraging covariance matrices to measure distances between data points and the distribution centre, ensured a robust assessment of outliers in a

multivariate context, enriching the integrity of subsequent analyses.

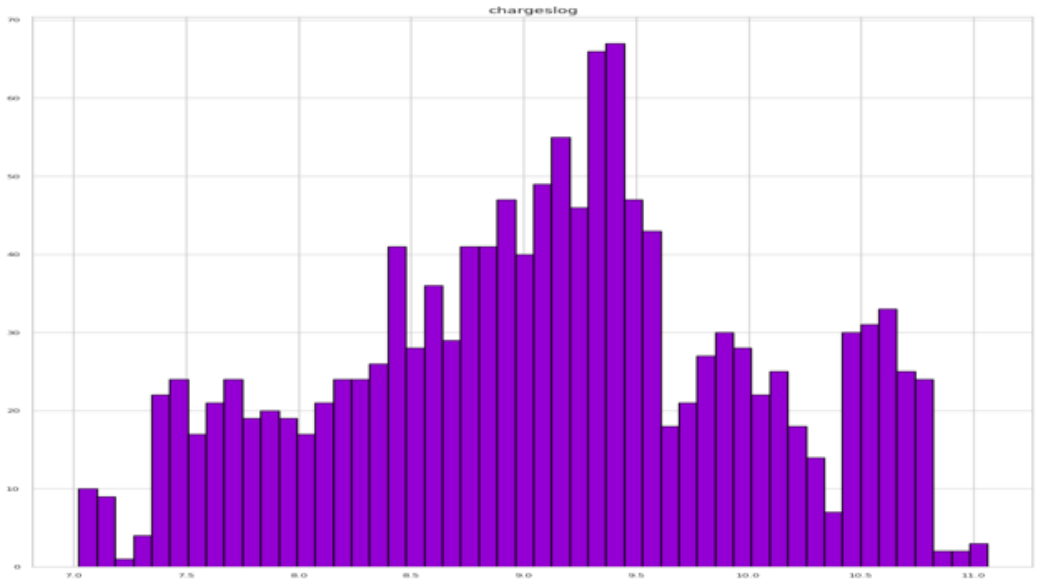



4. Categorical Features Scrutiny: A comprehensive analysis of categorical features ensued, involving count plots to visualize their distributions and impact. This meticulous examination aids in discerning the prevalence and significance of each categorical attribute, steering clear of predictors with overwhelming outcomes that might not substantially contribute to model training. The subsequent utilization of boxplots helped gauge the variation of insurance charges concerning each categorical feature, facilitating informed decisions on feature selection for modeling.

| | sex | smoker | region | charges |
|---|---|---|---|---|
| 0 | female | yes | southwest | 16,884.92 |
| 1 | male | no | southeast | 1,725.55 |
| 2 | male | no | southeast | 4,449.46 |
| 3 | male | no | northwest | 21,984.47 |
| 4 | male | no | northwest | 3,866.86 |
| ... | ... | ... | ... | ... |
| 1333 | male | no | northwest | 10,600.55 |
| 1334 | female | no | northeast | 2,205.98 |
| 1335 | female | no | southeast | 1,629.83 |
| 1336 | female | no | southwest | 2,007.94 |
| 1337 | female | yes | northwest | 29,141.36 |

[1338 rows x 4 columns]

Looking at the distribution of the numerical features right from the beginning, we can notice that "charges" is skewed as well. To help normalize this variable, a log transformation will be applied to "charges".




5. Feature Transformation and Normalization: Transforming categorical features into binary representations and applying log transformation on the target variable to achieve a normal distribution highlighted a strategic move towards feature engineering. This step aligns with enhancing the dataset's suitability for modeling, ensuring that categorical variables are appropriately represented and the target variable adheres to assumptions required by certain modeling techniques. Now the data changed into 1338x9 i.e., 9 features for our modelling.

| | age | bmi | children | charges | sex_male | smoker_yes | region_northwest | region_southeast | region_southwest |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 19 | 27.90 | 0 | 16,884.92 | 0 | 1 | 0 | 0 | 1 |
| **1** | 18 | 33.77 | 1 | 1,725.55 | 1 | 0 | 0 | 1 | 0 |
| **2** | 28 | 33.00 | 3 | 4,449.46 | 1 | 0 | 0 | 1 | 0 |
| **3** | 33 | 22.70 | 0 | 21,984.47 | 1 | 0 | 1 | 0 | 0 |
| **4** | 32 | 28.88 | 0 | 3,866.86 | 1 | 0 | 1 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1333** | 50 | 30.97 | 3 | 10,600.55 | 1 | 0 | 1 | 0 | 0 |
| **1334** | 18 | 31.92 | 0 | 2,205.98 | 0 | 0 | 0 | 0 | 0 |
| **1335** | 18 | 36.85 | 0 | 1,629.83 | 0 | 0 | 0 | 1 | 0 |
| **1336** | 21 | 25.80 | 0 | 2,007.94 | 0 | 0 | 0 | 0 | 1 |
| **1337** | 61 | 29.07 | 0 | 29,141.36 | 0 | 1 | 1 | 0 | 0 |

1338 rows x 9 columns

7. Split data into Train and Test sets: The dataset underwent a strategic split, yielding a training set of 1070 entries and a distinct test set comprising 268 entries, each containing eight essential features. This segregation ensured a clear demarcation between data for model training and subsequent evaluation, preserving the independence of the test data from the training process.

   Following the partitioning, attention shifted to transforming the data to meet the distinct requirements of machine learning algorithms. This process aimed not only at data transformation but also at selecting a representation that effectively revealed the intricate underlying structures within the predictive problem. By strategically choosing this representation, the algorithms were empowered to discern hidden patterns, optimizing their understanding and enhancing predictive performance within the project's resource constraints.

# Modeling Overview:

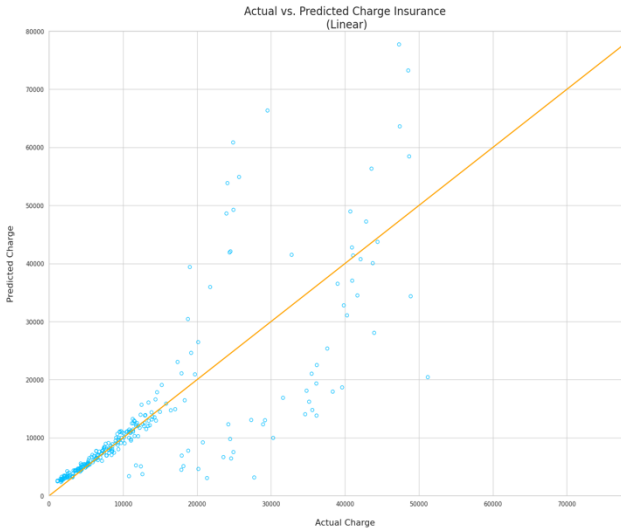
- Linear Regression
- Ridge Regression
- Lasso Regression
- Random Forest Regressor
- XGBoost Regressor

## Model Performance Evaluation: RMSE and R-squared Scores

I. RMSE (Root Mean Squared Error): RMSE quantifies the average difference between predicted and actual values. It's computed by taking the square root of the average squared differences between predictions and true values. Lower RMSE values indicate better model performance in predicting insurance charges.

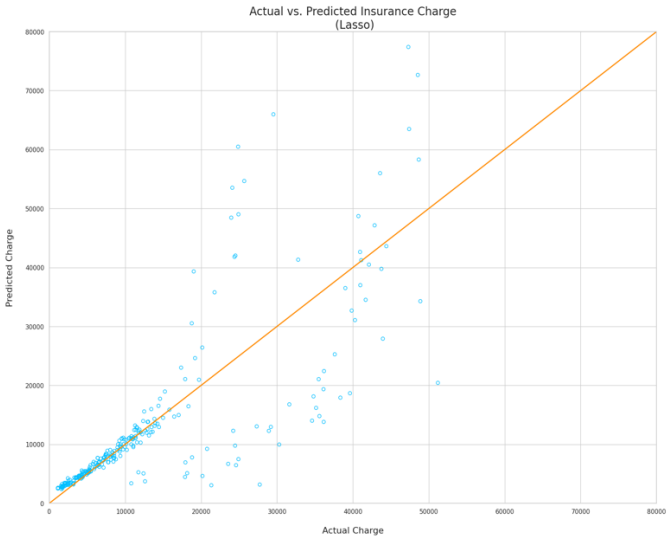
II. R-squared Score (Coefficient of Determination): R-squared measures the proportion of variance in the target variable explained by the model. A score closer to 1 implies a better fit of the model to the data, indicating how well the features explain the variability in insurance charges.

## Linear Regression:



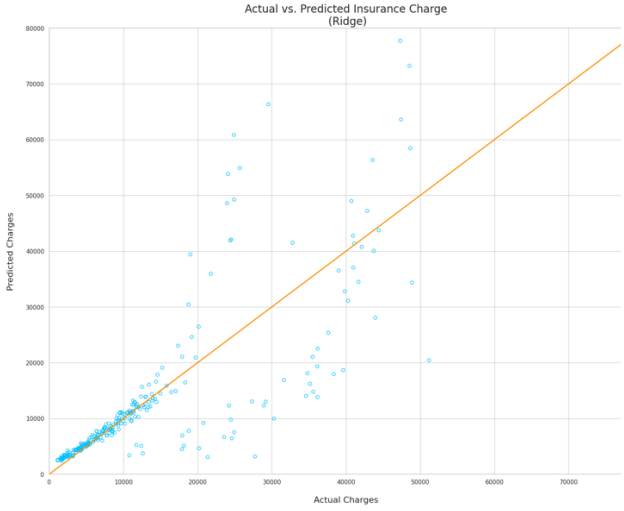

```
**** Regressor: Linear ****
R²: 0.7583042098055544
RMSE: 0.44680911669956613
```

## Lasso Regression:



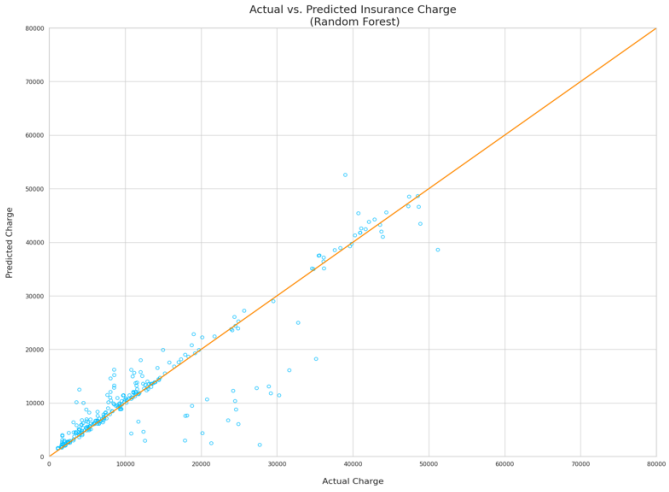

```
**** Regressor: Lasso ****
R²: 0.7584416589307031
RMSE: 0.44668205148280593
```

## Ridge Regression:



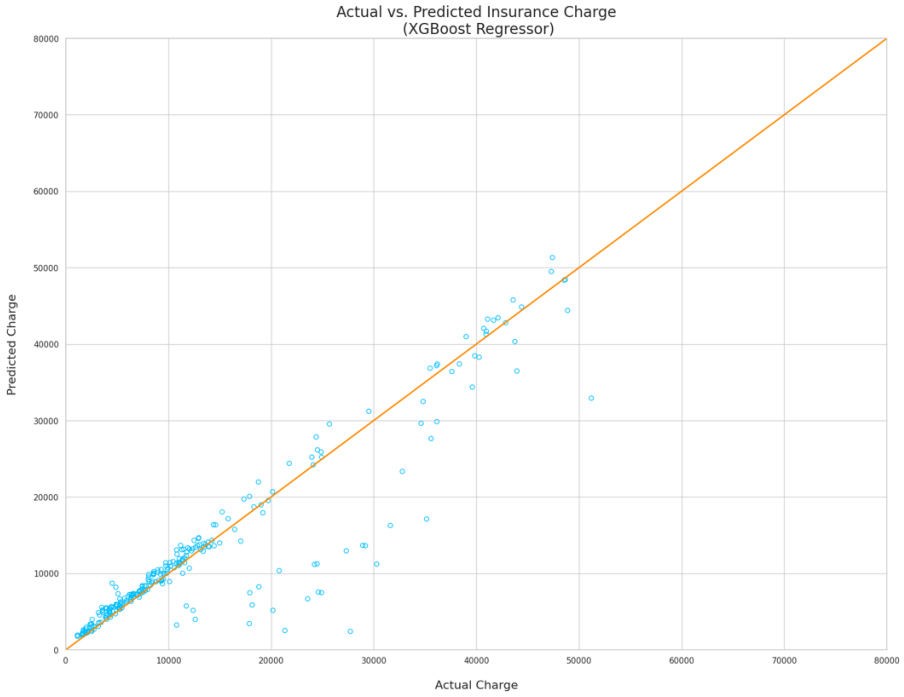

```
**** Regressor: Ridge ****
R²: 0.7583042175563077
RMSE: 0.44680910953537994
```

## Random Forest Regressor:




```
**** Regressor: Random Forest ****
R²: 0.7932162588191821
RMSE: 0.4132812259195818
```

## XGBoost Regression:




```
**** Regressor: XGBoost ****
R²: 0.8247486447824814
RMSE: 0.3804680349848538
```

## Results:

| Model | $R^2$ | RMSE |
|---|---|---|
| LinearRegression | 0.76 | 0.45 |
| Ridge | 0.76 | 0.45 |
| Lasso | 0.76 | 0.45 |
| RandomForestRegressor | 0.79 | 0.41 |
| XGBRegressor | 0.82 | 0.38 |

Among the suite of regression models including Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, and XGBoost Regressor, the XGBoost Regressor emerged as the frontrunner. It achieved the highest performance scores with an R-squared value of 0.82, indicating strong predictive capability in explaining the variance, while the RMSE (Root Mean Squared Error) stood at 0.38, signifying minimal average prediction error. This

model outperformed others in accurately capturing the relationships between variables, demonstrating its effectiveness in predicting the target variable within this context.

# Bayesian Methodology:

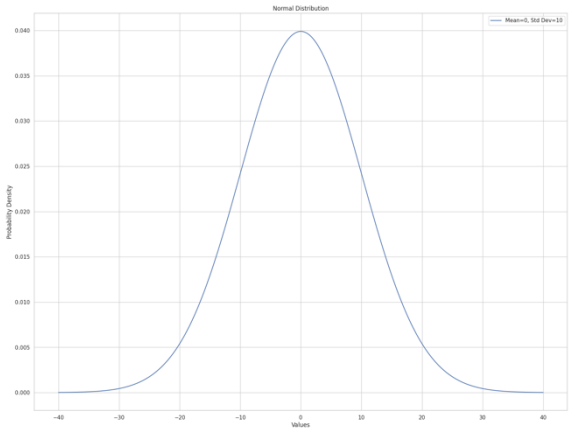
**MCMC**: The Bayesian approach, specifically using Markov Chain Monte Carlo (MCMC), is applied in this project for modeling health insurance costs. MCMC is used to estimate posterior distributions of model parameters(intercept & slope), allowing for uncertainty quantification and more robust predictions. The Bayesian approach contrasts with traditional frequentist methods by offering a comprehensive view of parameter uncertainty and incorporating prior knowledge.

# Prior Probability (Prior):

Before seeing any medical charges data, the prior distributions express our uncertainty about the parameters.

Here, the priors in our model are **P(Intercept),P(beta)**

These represent our initial beliefs about the intercept and slope (beta) parameters before observing the medical charges data. We are assuming normal distributions with mean 0 and standard deviation 10 as prior distributions for the intercept and beta parameters.

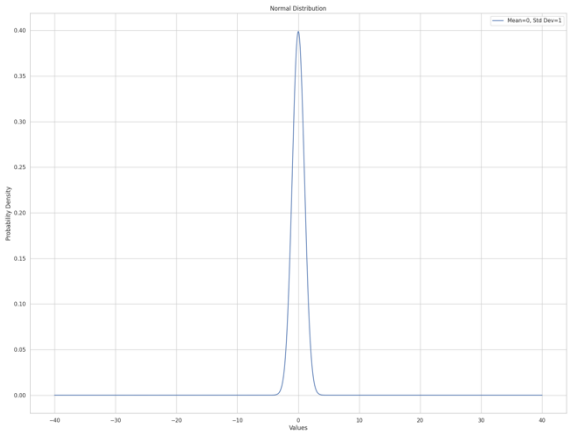



# Likelihood:

## P(charges|Intercept,beta,BMI)

It quantifies how likely observed charges are for different combinations of intercept, slope, and BMI.

This represents the probability of observing the medical charges data given specific values of the intercept, slope, and BMI. We are considering a normal likelihood assuming charges follow a normal distribution around the predicted values based on the linear model, with a standard deviation of 1.




# Posterior Probability (Posterior)

## *P*(Intercept,beta|charges,BMI)

After analysing the medical charges data, the posterior distributions represent our updated understanding of the intercept and slope values.

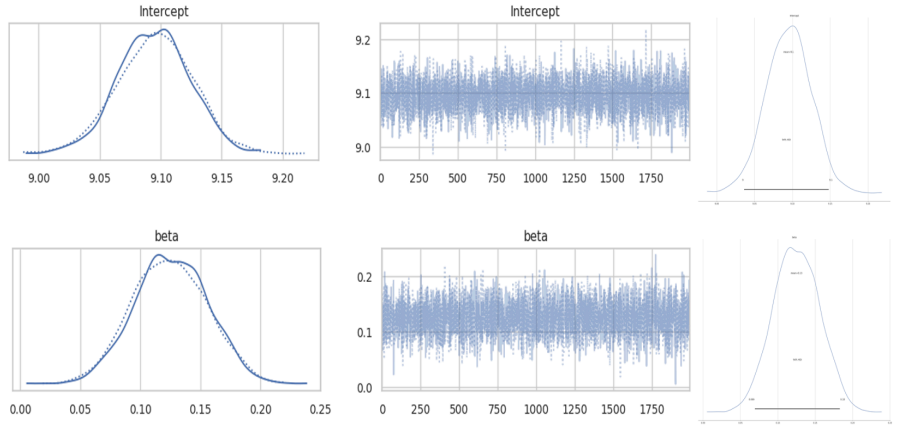
This is our updated belief about the intercept and beta parameters after considering the observed medical charges data and BMI. It's a combination of the prior beliefs and what the data tells us about the parameters.

$$P(\text{Intercept, beta}|\text{charges, BMI}) = \frac{P(\text{charges}|\text{Intercept,beta,BMI}) \times P(\text{Intercept}) \times P(\text{beta})}{P(\text{charges})}$$

- **Prior** distributions express initial uncertainty about the intercept and slope parameters.
- **Likelihood** calculates the probability of observing charges given specific intercept, slope, and BMI values.
- **Posterior** distributions represent updated beliefs about the intercept and slope parameters after observing medical charges data and BMI, combining prior knowledge with observed evidence.

# MCMC for Linear Regression Models:

The application of Bayesian linear regression with Markov Chain Monte Carlo (MCMC) revolved around leveraging the dataset's attributes. Among these variables, BMI stood out for its closer adherence to a normal distribution compared to others. This alignment with the normality assumption holds particular significance in linear regression models, especially within the Bayesian framework. Hence, the Bayesian linear regression analysis primarily centered on utilizing BMI as the key predictor for the target variable.



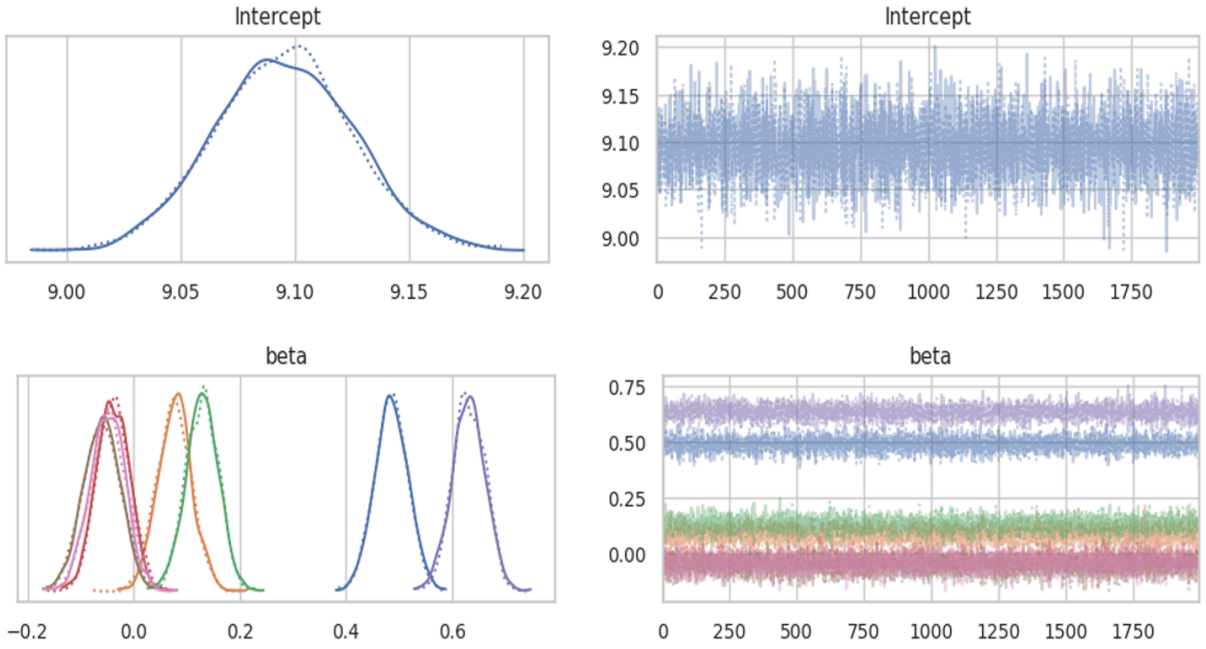

Mean squared error: 0.82

| | train_x | y_train | y_pred |
|---|---|---|---|
| **0** | 0.36 | 8.79 | 8.96 |
| **1** | -0.30 | 8.38 | 9.05 |
| **2** | -0.06 | 11.04 | 9.11 |
| **3** | -1.08 | 7.70 | 9.01 |
| **4** | -1.46 | 9.38 | 9.03 |
| **5** | 1.48 | 7.92 | 8.98 |
| **6** | 0.26 | 9.17 | 9.23 |
| **7** | -1.29 | 8.99 | 9.17 |
| **8** | -0.89 | 9.04 | 9.19 |
| **9** | 0.49 | 9.09 | 8.99 |

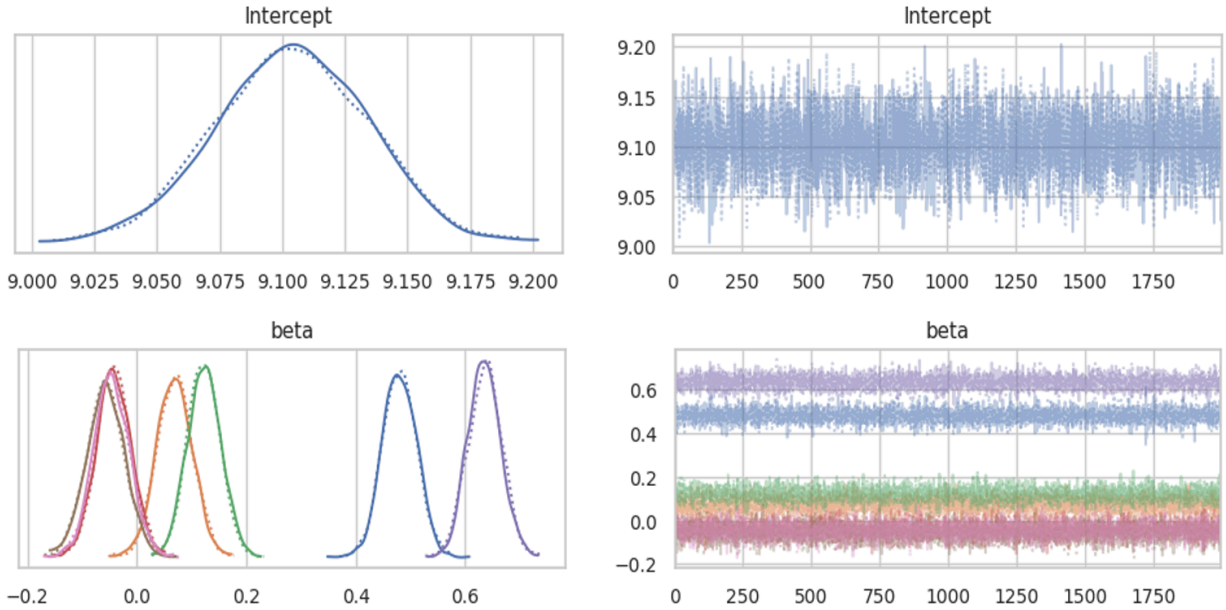
## MCMC for Multiple Linear Regression Models:

In this phase, I'm conducting multiple linear regression using a set of nine distinct features, including BMI, on the designated training dataset. While BMI is among these features, the analysis encompasses a range of additional variables that collectively contribute insights into predicting the target variable—insurance charges. This comprehensive examination aims to understand how these varied attributes collectively impact the outcome. By incorporating BMI alongside other factors, I'm exploring their individual and combined influences within the multiple linear regression framework, seeking to construct a robust predictive model that considers the interplay of these diverse features in predicting insurance charges.

## Results on Train(1070, 7) data:



Results on Test(268, 7) data:




```
Mean squared error: 0.17
Intercept    9.10
beta[0]      0.48
beta[1]      0.07
beta[2]      0.12
beta[3]     -0.04
beta[4]      0.64
beta[5]     -0.06
beta[6]     -0.05
Name: mean, dtype: float64
```

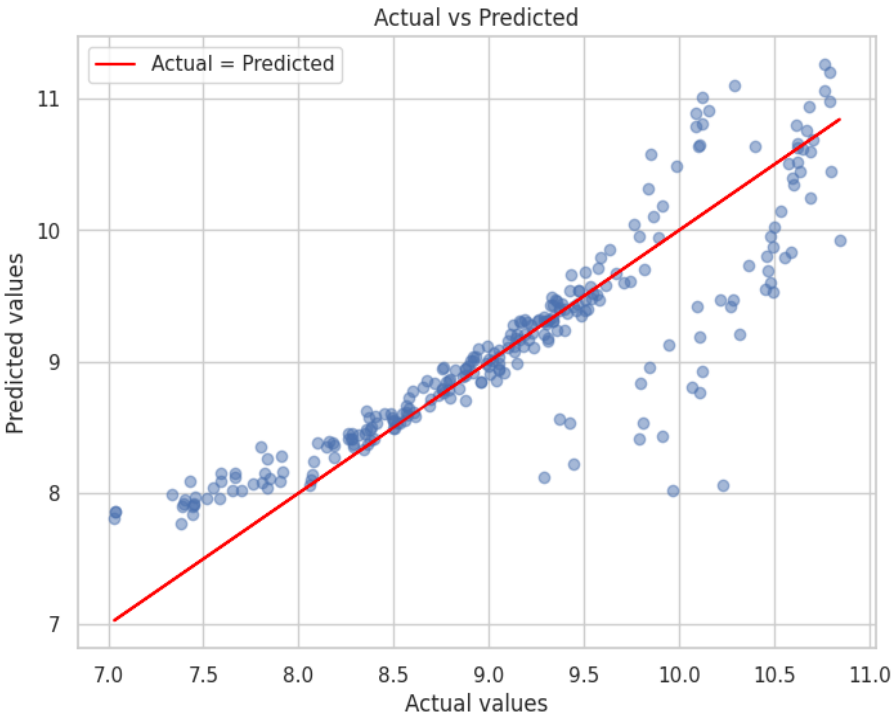
```
Coefficients:
 [ 9.10331276  0.48029169  0.06890278  0.12280198 -0.04199448  0.63518322
 -0.05770177 -0.04635644]
```

The Multiple Linear Regression equation is:

y_pred = beta[0]x1+ beta[1]x2+ beta[2]x3+ beta[3]x4+ beta[4]x5+ beta[5]x6+ beta[6]x7+Intercept

y_pred = 0.48(age)+ 0.07(bmi)+ 0.12(children)- 0.04(sex_male)+ 0.64 smoker_yes)
- 0.06(region_southeast)- 0.05(region_southwest)+9.10

**y_pred shape = (268, 2000)**




| | y_test | y_pred |
|---|---|---|
| 0 | 9.29 | 9.13 |
| 1 | 10.10 | 10.67 |
| 2 | 8.26 | 8.44 |
| 3 | 8.69 | 8.60 |
| 4 | 10.09 | 10.78 |
| 5 | 8.46 | 8.48 |
| 6 | 9.56 | 9.72 |
| 7 | 8.88 | 8.97 |
| 8 | 10.65 | 10.57 |
| 9 | 7.44 | 7.88 |

R-squared: 0.8940
RMSE: 0.4604

Without Bayesian:

| Model | $R^2$ | RMSE |
|---|---|---|
| LinearRegression | 0.76 | 0.45 |
| Ridge | 0.76 | 0.45 |
| Lasso | 0.76 | 0.45 |
| RandomForestRegressor | 0.79 | 0.41 |
| XGBRegressor | 0.82 | 0.38 |

With Bayesian:

| | Model | $R^2$ | RMSE |
|---|---|---|---|
| 0 | Bayesian Multi Linear Regression | 0.89 | 0.46 |

In contrast to various regression models explored, the Bayesian multiple linear regression exhibited superior performance, achieving the most favorable results. With an impressive R-squared value of 0.89, the model showcased exceptional ability in explaining the variance within the dataset. Additionally, boasting a low RMSE (Root Mean Squared Error) of 0.46, it demonstrated remarkable accuracy in predicting the target variable. This Bayesian approach surpassed other models, showcasing its efficacy in capturing intricate relationships among variables and making highly accurate predictions within this specific context.

# Conclusion:

The Bayesian linear regression may have a smaller mean squared error compared to the least squares linear regression due to the uncertainty estimates provided by the Bayesian model. In Bayesian linear regression, we estimate the posterior distribution of the model parameters given the data, which allows us to obtain a range of plausible values for the parameters, instead of a single point estimate as in least squares regression.

This uncertainty estimate can be particularly helpful in situations where we have limited data or when the data is noisy or has outliers. The uncertainty estimates can help to regularize the model and reduce overfitting. Additionally, Bayesian models allow for the incorporation of prior knowledge, which can be useful in situations where we have some prior knowledge about the relationship between the variables.

Therefore, the Bayesian approach may provide a more robust estimate of the parameters, resulting in better predictions and a smaller mean squared error compared to the least squares approach.

-\The End/-