

# Practice Project Overview



**Estimated Effort: 5 mins**

## Project Scenario

You have to perform data analytics on a medical insurance charges dataaset. This is a filtered and modified version of the [Medical Insurance Price Prediction](#) dataset, available under the [CC0 1.0 Universal License](#) on the [Kaggle](#) website.

## Parameters

The parameters used in the dataset are:

1. **Age**  
Age of the insured. Integer quantity.
2. **Gender**  
Gender of the insured. This parameter has been mapped to numerical values in the following way.

### **Gender Assigned Value**

Female 1  
Male 2

3. **BMI**  
Body Mass Index of the insured. Float value quantity.
4. **No\_of\_Children**

Number of children the insured person has. Integer quantity.

#### 5. **Smoker**

Whether the insured person is a smoker or not. This parameter has been mapped to numerical values in the following way.

<b>Smoker</b>	<b>Assigned Value</b>
Smoker	1
Non smoker	2

#### 6. **Region**

Which region of the USA does the insured belong to. This parameter has been mapped to numerical values in the following way.

<b>Region</b>	<b>Assigned Value</b>
Northwest	1
Northeast	2
Southwest	3
Southeast	4

#### 7. **Charges**

Charges for the insurance in USD. Floating value quantity.

## Objectives

In this project, you will:

- Load the data as a pandas dataframe
- Clean the data, taking care of the blank entries
- Run exploratory data analysis and identify the attributes that most affect the charges
- Develop single variable and multi variable Linear Regression models for predicting the charges
- Use Ridge regression to refine the performance of Linear regression models.

## Author(s)

[Abhishek Gagneja](#)

[Vicky Kuo](#)

## Changelog

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2023-09-16	0.1	Abhishek Gagneja	Initial Version Created
2023-09-19	0.2	Vicky Kuo	Reviewed and Revised

**Copyright © 2023 IBM Corporation. All rights reserved.**