# Health Insurance Cost Modeling

Kyle Offenloch

2025-08-30

### Data used in this Project:

Health Insurance dataset from kaggle.com

https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance

Includes information on attributes of 1338 insured individuals, including their age, sex, BMI, number of children, smoking habits, and region. It also includes their medical costs incurred.

### Goal of this project:

- Analyze data to infer effects of other variables on medical costs.
- Create generalized linear models using different assumptions to predict costs.
- Test and compare models.

## General Analysis

First, we break down the data set to look at the entire picture

```
set.seed(7997169)
split.ratio <- 2/3

library("ggplot2")
library("tidyr")
library("dplyr")
library("gridExtra")


#importing data
data <- read.csv("insurance.csv")
df<-data

train_indices<-sample(seq_len(nrow(data)), size = floor(split.ratio * nrow(data)))

train_data<-data[train_indices,]
test_data<-data[-train_indices,]

#making charts to view raw training data
agechart<-ggplot(train_data, aes(x = age)) +
  geom_histogram(binwidth = 1,fill = "steelblue", color = "white") +
  ggtitle("age Distribution") +
  theme_minimal()
sexchart<-ggplot(train_data, aes(x = "", fill=sex)) +
  geom_bar(width=1, color="blue") +
  coord_polar(theta = "y") +
  ggtitle("sex Distribution") +
  theme_void()
bmichart<-ggplot(train_data, aes(x=bmi)) +
  geom_histogram(binwidth=1, fill="steelblue", color="white") +
  ggtitle("bmi Distribution") +
  theme_minimal()
childrenchart<-ggplot(train_data, aes(x=children)) +
  geom_histogram(binwidth=1, fill="steelblue", color="white") +
  ggtitle("child Distribution") +
  theme_minimal()
smokerchart<-ggplot(train_data, aes(x = "", fill=smoker)) +
  geom_bar(width=1, color="blue") +
  coord_polar(theta = "y") +
  ggtitle("smoker Distribution") +
  theme_void()
regionchart<-ggplot(train_data, aes(x = "", fill=region)) +
  geom_bar(width=1, color="blue") +
  coord_polar(theta = "y") +
  ggtitle("region Distribution") +
  theme_void()
chargeschart<-ggplot(train_data, aes(x=charges)) +
  geom_histogram(bins = 50, fill="steelblue", color="white") +
  ggtitle("expenses Distribution") +
  theme_minimal()


#overview of all raw data
grid.arrange(agechart, sexchart, bmichart, childrenchart, smokerchart, regionchart, chargeschart)
```