R\_Car\_Accidents

Amogelang Alicia Seipone

2024-11-19

## Introduction

As part of my project, I decided to explore the concept of car insurance pricing using data analysis. Specifically, I wanted to look at how different factors, like age, gender, and accident history, affect the risk of getting into an accident and how these factors can influence the cost of car insurance premiums.

The goal was to create a synthetic dataset that mimics real-world data, perform some basic analysis on it, and simulate the pricing of car insurance based on risk scores.

## Data Generation

To get started, I generated synthetic data that includes: - **Age**: Ranging from 18 to 80 years, to simulate a variety of age groups. - **Gender**: A random assignment of male or female. - **Accident History**: A count of past accidents, ranging from 0 to 5. I used a random number generator to create these values, meaning the data doesn’t represent real individuals but gives us a feel for how these factors might influence risk.

# Load Required Libraries  
library(ggplot2) # For visualization

## Warning: package 'ggplot2' was built under R version 4.3.3

library(dplyr) # For data manipulation

## Warning: package 'dplyr' was built under R version 4.3.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# Generate Synthetic Data  
set.seed(123) # For reproducibility  
n <- 1000 # Number of samples  
  
synthetic\_data <- data.frame(  
 ID = 1:n,  
 Age = sample(18:80, n, replace = TRUE),  
 Gender = sample(c("Male", "Female"), n, replace = TRUE),  
 Accident\_History = sample(0:5, n, replace = TRUE)  
)

## Risk Scoring

The main part of the analysis involves calculating a Risk Score for each individual based on certain assumptions:

**Age:**

Under 25 years old: Higher risk score (80).

25 to 60 years old: Medium risk score (50).

Over 60 years old: Slightly higher risk score (70).

**Gender:** Females were assigned a risk score of 20, and males a risk score of 10.

**Accident History:** Each past accident adds 15 to the risk score.

# Calculate Risk Scores  
synthetic\_data <- synthetic\_data %>%  
 mutate(  
 Risk\_Score = case\_when(  
 Age < 25 ~ 80,  
 Age >= 25 & Age <= 60 ~ 50,  
 Age > 60 ~ 70,  
 TRUE ~ 0  
 ) + case\_when(  
 Gender == "Female" ~ 20,  
 Gender == "Male" ~ 10,  
 TRUE ~ 0  
 ) + Accident\_History \* 15  
 )

## Simulating Premiums

Once I had the risk scores, I calculated the insurance premium for each individual:

The premium is determined by multiplying the risk score by 5 and adding a base amount of 100.

# Simulate Premiums  
synthetic\_data <- synthetic\_data %>%  
 mutate(Premium = Risk\_Score \* 5 + 100)

This gives us an idea of how premiums might vary based on risk. For example, a person with a higher risk score (due to being younger or having more accidents) would have a higher premium.

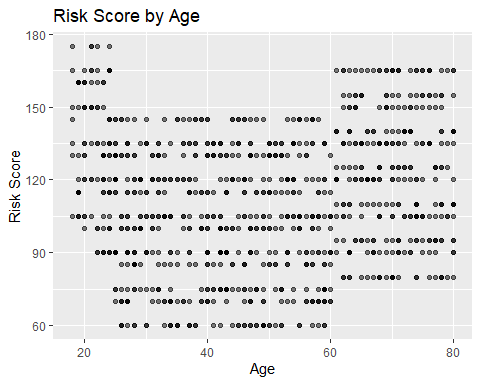
## Data Visualization

To visualize the data, I created two plots:

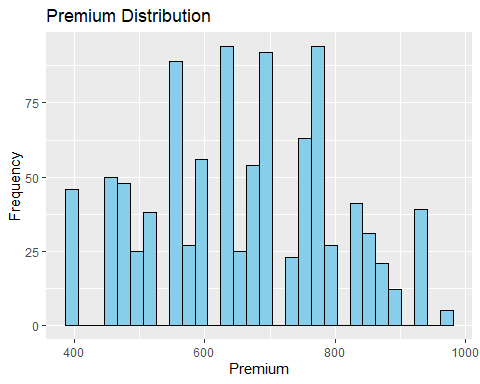
**Risk Score by Age:** A scatter plot that shows how risk scores vary across different age groups. As expected, younger individuals (under 25) generally have higher risk scores, while older individuals (over 60) also show a slight increase.

**Premium Distribution:** A histogram that shows how premiums are distributed across the dataset. Most individuals have a premium around the middle of the range, with fewer people at the extremes (very low or very high premiums).

# Visualize Risk Score by Age  
ggplot(synthetic\_data, aes(x = Age, y = Risk\_Score)) +  
 geom\_point(alpha = 0.5) +  
 labs(title = "Risk Score by Age", x = "Age", y = "Risk Score")



# Visualize Premium Distribution  
ggplot(synthetic\_data, aes(x = Premium)) +  
 geom\_histogram(bins = 30, fill = "skyblue", color = "black") +  
 labs(title = "Premium Distribution", x = "Premium", y = "Frequency")



## Results

**The data shows clear patterns:**

Younger people (especially under 25) tend to have higher risk scores and, therefore, higher premiums. People with more accidents in their history also see a rise in premiums. The gender difference in the risk score is minimal but is included here to show how you could incorporate different factors into the model.

# Export Results  
write.csv(synthetic\_data, "synthetic\_insurance\_data.csv", row.names = FALSE)  
  
# Save plots as images  
ggsave("risk\_score\_by\_age.png", width = 8, height = 5)  
ggsave("premium\_distribution.png", width = 8, height = 5)