Synthetic Insurance Analysis

Anshu Kumar & Sonu Kumar

2025-04-27

##Load Necessary Libraries

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)  
library(readr)  
library(MASS)

##   
## Attaching package: 'MASS'  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

library(ggplot2)  
library(GGally)

## Warning: package 'GGally' was built under R version 4.4.3

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

*Firstly we take important Libraries so we can perform our next task*

# 1. Data Understanding

## Load Dataset

insurance\_data = read\_csv("C:/Users/LENOVO/Downloads/insurance.zip")

## Rows: 10000 Columns: 27  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (6): Marital\_Status, Prior\_Insurance, Claims\_Severity, Policy\_Type, Sou...  
## dbl (21): Age, Is\_Senior, Married\_Premium\_Discount, Prior\_Insurance\_Premium\_...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(insurance\_data)

*We started by loading the insurance dataset using read\_csv() and viewed it with View() to familiarize ourselves with the data.*

##Look the structure & Summarization of dataset

glimpse(insurance\_data)

## Rows: 10,000  
## Columns: 27  
## $ Age <dbl> 47, 37, 49, 62, 36, 36, 63, 51, 32,…  
## $ Is\_Senior <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0,…  
## $ Marital\_Status <chr> "Married", "Married", "Married", "M…  
## $ Married\_Premium\_Discount <dbl> 86, 86, 86, 86, 0, 86, 86, 0, 86, 0…  
## $ Prior\_Insurance <chr> "1-5 years", "1-5 years", "1-5 year…  
## $ Prior\_Insurance\_Premium\_Adjustment <dbl> 50, 50, 50, 0, 0, 0, 50, 100, 0, 0,…  
## $ Claims\_Frequency <dbl> 0, 0, 1, 1, 2, 0, 0, 0, 0, 1, 1, 1,…  
## $ Claims\_Severity <chr> "Low", "Low", "Low", "Low", "Low", …  
## $ Claims\_Adjustment <dbl> 0, 0, 50, 50, 100, 0, 0, 0, 0, 200,…  
## $ Policy\_Type <chr> "Full Coverage", "Full Coverage", "…  
## $ Policy\_Adjustment <dbl> 0, 0, 0, 0, 0, -200, 0, 0, -200, 0,…  
## $ Premium\_Amount <dbl> 2286, 2336, 2386, 2336, 2350, 1936,…  
## $ Safe\_Driver\_Discount <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0,…  
## $ Multi\_Policy\_Discount <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0,…  
## $ Bundling\_Discount <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,…  
## $ Total\_Discounts <dbl> 0, 0, 0, 0, 0, 50, 50, 50, 50, 50, …  
## $ Source\_of\_Lead <chr> "Agent", "Online", "Online", "Onlin…  
## $ Time\_Since\_First\_Contact <dbl> 10, 22, 28, 4, 14, 13, 2, 1, 16, 27…  
## $ Conversion\_Status <dbl> 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1,…  
## $ Website\_Visits <dbl> 5, 5, 4, 6, 8, 4, 5, 3, 5, 5, 3, 3,…  
## $ Inquiries <dbl> 1, 1, 4, 2, 4, 1, 1, 0, 1, 3, 2, 1,…  
## $ Quotes\_Requested <dbl> 2, 2, 1, 2, 2, 1, 2, 2, 3, 2, 1, 3,…  
## $ Time\_to\_Conversion <dbl> 99, 99, 99, 2, 10, 7, 1, 99, 3, 99,…  
## $ Credit\_Score <dbl> 704, 726, 772, 809, 662, 729, 795, …  
## $ Premium\_Adjustment\_Credit <dbl> -50, -50, -50, -50, 50, -50, -50, 5…  
## $ Region <chr> "Suburban", "Urban", "Urban", "Urba…  
## $ Premium\_Adjustment\_Region <dbl> 50, 100, 100, 100, 50, 0, 100, 50, …

summary(insurance\_data)

## Age Is\_Senior Marital\_Status Married\_Premium\_Discount  
## Min. :18.00 Min. :0.0000 Length:10000 Min. : 0.00   
## 1st Qu.:29.00 1st Qu.:0.0000 Class :character 1st Qu.: 0.00   
## Median :39.00 Median :0.0000 Mode :character Median : 0.00   
## Mean :39.99 Mean :0.1593 Mean :42.13   
## 3rd Qu.:50.00 3rd Qu.:0.0000 3rd Qu.:86.00   
## Max. :90.00 Max. :1.0000 Max. :86.00   
## Prior\_Insurance Prior\_Insurance\_Premium\_Adjustment Claims\_Frequency  
## Length:10000 Min. : 0.00 Min. :0.0000   
## Class :character 1st Qu.: 0.00 1st Qu.:0.0000   
## Mode :character Median : 50.00 Median :0.0000   
## Mean : 47.62 Mean :0.4972   
## 3rd Qu.: 50.00 3rd Qu.:1.0000   
## Max. :100.00 Max. :5.0000   
## Claims\_Severity Claims\_Adjustment Policy\_Type Policy\_Adjustment  
## Length:10000 Min. : 0.00 Length:10000 Min. :-200.00   
## Class :character 1st Qu.: 0.00 Class :character 1st Qu.:-200.00   
## Mode :character Median : 0.00 Mode :character Median : 0.00   
## Mean : 36.78 Mean : -79.86   
## 3rd Qu.: 50.00 3rd Qu.: 0.00   
## Max. :800.00 Max. : 0.00   
## Premium\_Amount Safe\_Driver\_Discount Multi\_Policy\_Discount Bundling\_Discount  
## Min. :1800 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:2100 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :2236 Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :2220 Mean :0.1999 Mean :0.3051 Mean :0.0972   
## 3rd Qu.:2336 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :2936 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## Total\_Discounts Source\_of\_Lead Time\_Since\_First\_Contact Conversion\_Status  
## Min. : 0.00 Length:10000 Min. : 1.00 Min. :0.0000   
## 1st Qu.: 0.00 Class :character 1st Qu.: 8.00 1st Qu.:0.0000   
## Median : 50.00 Mode :character Median :16.00 Median :1.0000   
## Mean : 30.11 Mean :15.48 Mean :0.5767   
## 3rd Qu.: 50.00 3rd Qu.:23.00 3rd Qu.:1.0000   
## Max. :150.00 Max. :30.00 Max. :1.0000   
## Website\_Visits Inquiries Quotes\_Requested Time\_to\_Conversion  
## Min. : 0.000 Min. :0.000 Min. :1.000 Min. : 1.00   
## 1st Qu.: 3.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.: 6.00   
## Median : 5.000 Median :2.000 Median :2.000 Median :12.00   
## Mean : 5.023 Mean :1.997 Mean :1.997 Mean :46.07   
## 3rd Qu.: 6.000 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.:99.00   
## Max. :16.000 Max. :9.000 Max. :3.000 Max. :99.00   
## Credit\_Score Premium\_Adjustment\_Credit Region   
## Min. :530.0 Min. :-50.00 Length:10000   
## 1st Qu.:681.0 1st Qu.:-50.00 Class :character   
## Median :715.0 Median :-50.00 Mode :character   
## Mean :714.3 Mean :-11.32   
## 3rd Qu.:748.0 3rd Qu.: 50.00   
## Max. :850.0 Max. : 50.00   
## Premium\_Adjustment\_Region  
## Min. : 0.00   
## 1st Qu.: 50.00   
## Median : 50.00   
## Mean : 64.33   
## 3rd Qu.:100.00   
## Max. :100.00

*We then used glimpse() to get a quick overview of the dataset’s structure and summary() to understand the statistical properties of each variable.*

# 2. Data Cleaning

## Check for Missing Values

sum(is.na(insurance\_data))

## [1] 0

*We checked for any missing values using sum(is.na()) and found whether our data was complete.*

##Remove Duplicate Records

insurance\_data = distinct(insurance\_data)  
distinct(insurance\_data)

## # A tibble: 10,000 × 27  
## Age Is\_Senior Marital\_Status Married\_Premium\_Discount Prior\_Insurance  
## <dbl> <dbl> <chr> <dbl> <chr>   
## 1 47 0 Married 86 1-5 years   
## 2 37 0 Married 86 1-5 years   
## 3 49 0 Married 86 1-5 years   
## 4 62 1 Married 86 >5 years   
## 5 36 0 Single 0 >5 years   
## 6 36 0 Married 86 >5 years   
## 7 63 1 Married 86 1-5 years   
## 8 51 0 Single 0 <1 year   
## 9 32 0 Married 86 >5 years   
## 10 48 0 Single 0 >5 years   
## # ℹ 9,990 more rows  
## # ℹ 22 more variables: Prior\_Insurance\_Premium\_Adjustment <dbl>,  
## # Claims\_Frequency <dbl>, Claims\_Severity <chr>, Claims\_Adjustment <dbl>,  
## # Policy\_Type <chr>, Policy\_Adjustment <dbl>, Premium\_Amount <dbl>,  
## # Safe\_Driver\_Discount <dbl>, Multi\_Policy\_Discount <dbl>,  
## # Bundling\_Discount <dbl>, Total\_Discounts <dbl>, Source\_of\_Lead <chr>,  
## # Time\_Since\_First\_Contact <dbl>, Conversion\_Status <dbl>, …

*We removed any duplicate records by applying distinct(), ensuring that our dataset only contains unique observations.*

# 3. Descriptive Analysis

## Average age of policyholders

average\_age = mean(insurance\_data$Age)  
print(paste("Average Age:", average\_age))

## [1] "Average Age: 39.9917"

*We calculated the average age of the policyholders to understand the typical customer profile.*

##Proportion of senior citizens

proportion\_senior = mean(insurance\_data$Is\_Senior)  
print(paste("Proportion of Senior Citizens:", proportion\_senior))

## [1] "Proportion of Senior Citizens: 0.1593"

*We found the proportion of senior citizens by taking the mean of the Is\_Senior column.*

##Distribution of marital status

marital\_status\_dist = insurance\_data %>%  
 group\_by(Marital\_Status) %>%  
 summarise(count = n())  
print(marital\_status\_dist)

## # A tibble: 4 × 2  
## Marital\_Status count  
## <chr> <int>  
## 1 Divorced 920  
## 2 Married 4899  
## 3 Single 3259  
## 4 Widowed 922

*We examined the distribution of marital status by grouping and counting the number of policyholders in each category.*

##Average premium amount

average\_premium = mean(insurance\_data$Premium\_Amount)  
print(paste("Average Premium Amount:", average\_premium))

## [1] "Average Premium Amount: 2219.5714"

*We computed the average premium amount to get a sense of the typical insurance cost.*

##Average premium amount for each region

average\_premium\_region = insurance\_data %>%  
 group\_by(Region) %>%  
 summarise(average\_premium = mean(Premium\_Amount))  
print(average\_premium\_region)

## # A tibble: 3 × 2  
## Region average\_premium  
## <chr> <dbl>  
## 1 Rural 2157.  
## 2 Suburban 2202.  
## 3 Urban 2256.

*We analyzed the average premium amount across different regions to see if there were any geographical differences.*

##Distribution of claims frequency

claims\_frequency\_dist = insurance\_data %>%  
 group\_by(Claims\_Frequency) %>%  
 summarise(count = n())  
print(claims\_frequency\_dist)

## # A tibble: 6 × 2  
## Claims\_Frequency count  
## <dbl> <int>  
## 1 0 6126  
## 2 1 2965  
## 3 2 745  
## 4 3 141  
## 5 4 21  
## 6 5 2

*We explored the distribution of claims frequency by counting occurrences in category.*

##Distribution of claims severity

claims\_severity\_dist = insurance\_data %>%  
 group\_by(Claims\_Severity) %>%  
 summarise(count = n())  
print(claims\_severity\_dist)

## # A tibble: 3 × 2  
## Claims\_Severity count  
## <chr> <int>  
## 1 High 959  
## 2 Low 7003  
## 3 Medium 2038

*We explored the distribution of claims severity by counting occurrences in category.*

##Average age by marital status

average\_age\_marital\_status = insurance\_data %>%  
 group\_by(Marital\_Status) %>%  
 summarise(average\_age = mean(Age))  
print(average\_age\_marital\_status)

## # A tibble: 4 × 2  
## Marital\_Status average\_age  
## <chr> <dbl>  
## 1 Divorced 42.3  
## 2 Married 39.8  
## 3 Single 39.1  
## 4 Widowed 41.8

*We calculated the average age by marital status to check if age varies between different marital statuses.*

##Most common claims severity

common\_claims\_severity = insurance\_data %>%  
 group\_by(Claims\_Severity) %>%  
 summarise(count = n()) %>%  
 arrange(desc(count)) %>%  
 slice(1)  
print(common\_claims\_severity)

## # A tibble: 1 × 2  
## Claims\_Severity count  
## <chr> <int>  
## 1 Low 7003

*We identified the most common claims severity level among the policyholders.*

# 4. Correlation and Relationship Analysis

## Correlation between Age and Premium Amount

correlation\_age\_premium = cor(insurance\_data$Age, insurance\_data$Premium\_Amount)  
print(paste("Correlation between Age and Premium Amount:", correlation\_age\_premium))

## [1] "Correlation between Age and Premium Amount: -0.0295406633794125"

*We calculated the correlation between age and premium amount to understand if age impacts how much premium a person pays.*

##Correlation between Prior Insurance Premium Adjustment and Current Premium Amount

correlation\_prior\_current\_premium = cor(  
 insurance\_data$Prior\_Insurance\_Premium\_Adjustment,  
 insurance\_data$Premium\_Amount)  
print(paste("Correlation between Prior Insurance Premium Adjustment and Current Premium Amount:", correlation\_prior\_current\_premium))

## [1] "Correlation between Prior Insurance Premium Adjustment and Current Premium Amount: 0.234540830168632"

*We measured the correlation between prior insurance premium adjustments and current premium amounts to see if there was any relationship.*

##Effect of Prior Insurance on Claims Frequency and Severity

prior\_insurance\_effect = insurance\_data %>%  
 group\_by(Prior\_Insurance) %>%  
 summarise(  
 average\_claims\_frequency = mean(Claims\_Frequency),  
 average\_claims\_severity = mean(as.numeric(Claims\_Severity))  
 )

## Warning: There were 3 warnings in `summarise()`.  
## The first warning was:  
## ℹ In argument: `average\_claims\_severity = mean(as.numeric(Claims\_Severity))`.  
## ℹ In group 1: `Prior\_Insurance = "1-5 years"`.  
## Caused by warning in `mean()`:  
## ! NAs introduced by coercion  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 2 remaining warnings.

print(prior\_insurance\_effect)

## # A tibble: 3 × 3  
## Prior\_Insurance average\_claims\_frequency average\_claims\_severity  
## <chr> <dbl> <dbl>  
## 1 1-5 years 0.492 NA  
## 2 <1 year 0.493 NA  
## 3 >5 years 0.512 NA

*We analyzed how prior insurance affects claims frequency and severity by grouping and summarizing the data.*

##Relationship between Policy Type and Premium Amount

policy\_type\_premium = insurance\_data %>%  
 group\_by(Policy\_Type) %>%  
 summarise(average\_premium = mean(Premium\_Amount))  
print(policy\_type\_premium)

## # A tibble: 2 × 2  
## Policy\_Type average\_premium  
## <chr> <dbl>  
## 1 Full Coverage 2300.  
## 2 Liability-Only 2099.

*We checked how the type of policy affects the average premium by grouping data by Policy\_Type.*

##Categorizing Credit Score Ranges

insurance\_data = insurance\_data %>%  
 mutate(Credit\_Score\_Range = case\_when(  
 Credit\_Score >= 750 ~ "Excellent (750+)",  
 Credit\_Score >= 700 & Credit\_Score < 750 ~ "Good (700-749)",  
 Credit\_Score >= 650 & Credit\_Score < 700 ~ "Fair (650-699)",  
 Credit\_Score < 650 ~ "Poor (<650)"  
 ))

*We categorized credit scores into meaningful ranges (Excellent, Good, Fair, Poor) to better segment the customers.* *Its create One another Column in dataset*

##Average premium amount by Credit Score Range

average\_premium\_credit\_score = insurance\_data %>%  
 group\_by(Credit\_Score\_Range) %>%  
 summarise(average\_premium = mean(Premium\_Amount))  
print(average\_premium\_credit\_score)

## # A tibble: 4 × 2  
## Credit\_Score\_Range average\_premium  
## <chr> <dbl>  
## 1 Excellent (750+) 2184.  
## 2 Fair (650-699) 2279.  
## 3 Good (700-749) 2182.  
## 4 Poor (<650) 2284.

*We calculated the average premium amount for each credit score range to see if creditworthiness affects premiums.*

##Premium differences by Marital Status

premium\_marital\_status = insurance\_data %>%  
 group\_by(Marital\_Status) %>%  
 summarise(average\_premium = mean(Premium\_Amount))  
print(premium\_marital\_status)

## # A tibble: 4 × 2  
## Marital\_Status average\_premium  
## <chr> <dbl>  
## 1 Divorced 2173.  
## 2 Married 2264.  
## 3 Single 2179.  
## 4 Widowed 2175.

*We compared premium differences across different marital statuses.*

##Claims Frequency differences by Policy Type

claims\_frequency\_policy = insurance\_data %>%  
 group\_by(Policy\_Type) %>%  
 summarise(average\_claims\_frequency = mean(Claims\_Frequency))  
print(claims\_frequency\_policy)

## # A tibble: 2 × 2  
## Policy\_Type average\_claims\_frequency  
## <chr> <dbl>  
## 1 Full Coverage 0.497  
## 2 Liability-Only 0.497

*We examined how claims frequency varies across different policy types.*

##Claims Adjustment by Policy Type

claims\_adjustment\_policy = insurance\_data %>%  
 group\_by(Policy\_Type) %>%  
 summarise(average\_claims\_adjustment = mean(Claims\_Adjustment))  
print(claims\_adjustment\_policy)

## # A tibble: 2 × 2  
## Policy\_Type average\_claims\_adjustment  
## <chr> <dbl>  
## 1 Full Coverage 36.4  
## 2 Liability-Only 37.3

*We studied how claims adjustment differs between various policy types.*

##Effect of Prior Insurance Premium Adjustment on Current Premium Amount

effect\_prior\_insurance = insurance\_data %>%  
 group\_by(Prior\_Insurance\_Premium\_Adjustment) %>%  
 summarise(average\_current\_premium = mean(Premium\_Amount))  
print(effect\_prior\_insurance)

## # A tibble: 3 × 2  
## Prior\_Insurance\_Premium\_Adjustment average\_current\_premium  
## <dbl> <dbl>  
## 1 0 2174.  
## 2 50 2220.  
## 3 100 2276.

*We studied how claims adjustment differs between various policy types.*

##Effect of Claims Adjustment on Premium Amount

effect\_claims\_adjustment = insurance\_data %>%  
 group\_by(Claims\_Adjustment) %>%  
 summarise(average\_premium\_amount = mean(Premium\_Amount))  
print(effect\_claims\_adjustment)

## # A tibble: 10 × 2  
## Claims\_Adjustment average\_premium\_amount  
## <dbl> <dbl>  
## 1 0 2183.  
## 2 50 2229.  
## 3 100 2289.  
## 4 150 2334.  
## 5 200 2379.  
## 6 250 2443   
## 7 300 2484.  
## 8 400 2560.  
## 9 600 2802.  
## 10 800 2900

*We also explored how claims adjustment levels are related to premium amounts.*

# 5. Data Regression

##Predict Premium Amount based on Age and Credit Score

Predict\_Premium\_Amount = lm(Premium\_Amount ~ Age + Credit\_Score, data = insurance\_data)  
summary(Predict\_Premium\_Amount)

##   
## Call:  
## lm(formula = Premium\_Amount ~ Age + Credit\_Score, data = insurance\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -426.41 -105.61 8.95 102.44 758.00   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2767.44236 21.07567 131.310 <2e-16 \*\*\*  
## Age -0.30694 0.10229 -3.001 0.0027 \*\*   
## Credit\_Score -0.74987 0.02889 -25.958 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 143.7 on 9997 degrees of freedom  
## Multiple R-squared: 0.06396, Adjusted R-squared: 0.06378   
## F-statistic: 341.6 on 2 and 9997 DF, p-value: < 2.2e-16

*We built a linear regression model to predict premium amount based on age and credit score.This helped us understand the impact of these factors on the premium.* *summary() tells you how well the model fits, coefficients, R-squared, etc.*

##Predict Claims Frequency based on Age, Credit Score, and Prior Insurance

Predict\_Claims\_Frequency = lm(Claims\_Frequency ~ Age + Credit\_Score + Prior\_Insurance, data = insurance\_data)  
summary(Predict\_Claims\_Frequency)

##   
## Call:  
## lm(formula = Claims\_Frequency ~ Age + Credit\_Score + Prior\_Insurance,   
## data = insurance\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5240 -0.4973 -0.4882 0.5041 4.5204   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.855e-01 1.059e-01 4.586 4.57e-06 \*\*\*  
## Age -3.698e-04 5.137e-04 -0.720 0.472   
## Credit\_Score 3.103e-05 1.440e-04 0.216 0.829   
## Prior\_Insurance>5 years 1.991e-02 2.103e-02 0.947 0.344   
## Prior\_Insurance1-5 years -1.628e-03 1.839e-02 -0.089 0.929   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7162 on 9995 degrees of freedom  
## Multiple R-squared: 0.000202, Adjusted R-squared: -0.0001981   
## F-statistic: 0.5049 on 4 and 9995 DF, p-value: 0.7322

*We developed another linear regression model to predict claims frequency using age, credit score, and prior insurance status.* *This model allowed us to study the factors affecting how often customers make claims.*

# 6. ANOVA Test

##Premium Amount difference across Marital Status

Premium\_Amount\_Across\_Marital\_Status = aov(Premium\_Amount ~ Marital\_Status, data = insurance\_data)  
summary(Premium\_Amount\_Across\_Marital\_Status)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Marital\_Status 3 18779068 6259689 310.1 <2e-16 \*\*\*  
## Residuals 9996 201784139 20186   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*We performed a one-way ANOVA to determine whether the average Premium Amount differs significantly across different Marital Status groups (Single, Married, Divorced, etc.).*

*Null Hypothesis (H₀): There is no difference in Premium Amount between marital status groups.* *Alternative Hypothesis (H₁): At least one group has a different mean Premium Amount* *If the p-value is less than 0.05, we reject the null hypothesis, meaning that marital status does impact the Premium Amount.* *Conclusion Example:* *If p-value = 0.002 → “There is a significant difference in Premium Amount between marital status groups.”*

##Claims Frequency difference across Policy Type

Claims\_Frequency\_across\_Policy\_Type = aov(Claims\_Frequency ~ Policy\_Type, data = insurance\_data)  
summary(Claims\_Frequency\_across\_Policy\_Type)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Policy\_Type 1 0 0.0000 0 0.993  
## Residuals 9998 5128 0.5129

*We conducted a one-way ANOVA to check if the Claims Frequency significantly varies across different Policy Types.* *Null Hypothesis (H₀): The mean Claims Frequency is the same for all Policy Types.* *Alternative Hypothesis (H₁): At least one Policy Type has a different mean Claims Frequency.* *If the p-value is less than 0.05, we conclude that Policy Type affects how frequently claims are made.* *Conclusion Example:* *If p-value = 0.01 → “There is a significant difference in Claims Frequency between Policy Types, indicating that certain policy types are riskier.”*

##Premium Amount difference across Regions

Premium\_Amount\_difference\_across\_Regions = aov(Premium\_Amount ~ Region, data = insurance\_data)  
summary(Premium\_Amount\_difference\_across\_Regions)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Region 2 15632513 7816257 381.3 <2e-16 \*\*\*  
## Residuals 9997 204930694 20499   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

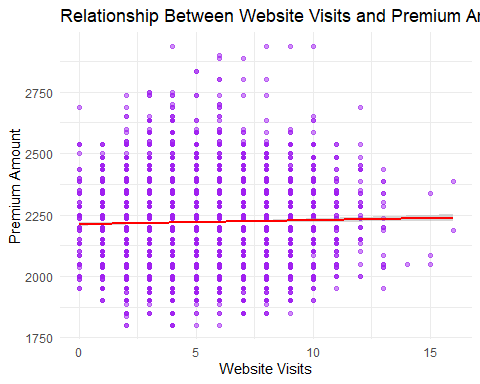
*We performed a one-way ANOVA to investigate whether the Premium Amount significantly differs across different Regions.* *Null Hypothesis (H₀): The mean Premium Amount is the same across all regions.* *Alternative Hypothesis (H₁): At least one region has a different mean Premium Amount.* *A p-value less than 0.05 would suggest that the Premium Amount varies significantly by Region.* *Conclusion Example:* *If p-value = 0.04 → “Premiums differ significantly across regions, suggesting that geographic factors influence pricing.”*

# 7. Data Visualization

##Website Visits vs Premium Amount

ggplot(insurance\_data, aes(x = Website\_Visits, y = Premium\_Amount)) +  
 geom\_point(alpha = 0.5, color = "purple") +  
 geom\_smooth(method = "lm", color = "red", se = TRUE) +  
 labs(title = "Relationship Between Website Visits and Premium Amount",  
 x = "Website Visits", y = "Premium Amount") +  
 theme\_minimal()

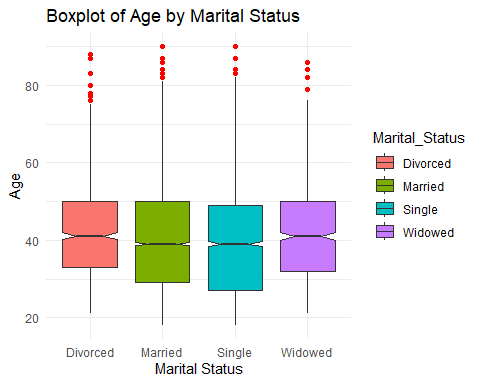
## `geom\_smooth()` using formula = 'y ~ x'



*We visualized the relationship between website visits and premium amount using a scatter plot with a linear trend line.*

# Boxplot of Age by Marital Status

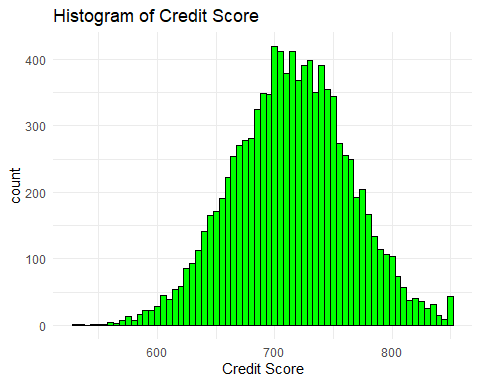
ggplot(insurance\_data, aes(x = Marital\_Status, y = Age, fill = Marital\_Status)) +  
 geom\_boxplot(outlier.color = "red", notch = TRUE) +  
 labs(title = "Boxplot of Age by Marital Status", x = "Marital Status", y = "Age") +  
 theme\_minimal()



*We created a boxplot to compare the age distributions across different marital statuses.*

# Histogram of Credit Score

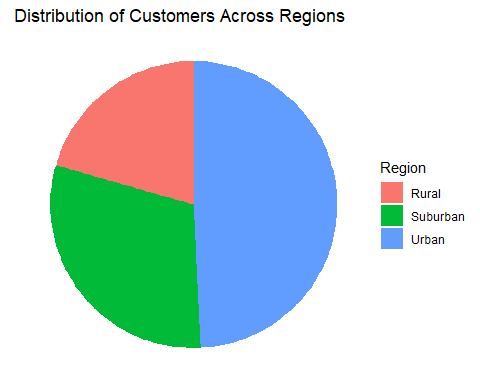
ggplot(insurance\_data, aes(x = Credit\_Score)) +  
 geom\_histogram(binwidth = 5, fill = "green", color = "black") +  
 labs(title = "Histogram of Credit Score", x = "Credit Score") +  
 theme\_minimal()



*We plotted a histogram to explore the distribution of credit scores among policyholders.*

#Customer Distribution across Regions

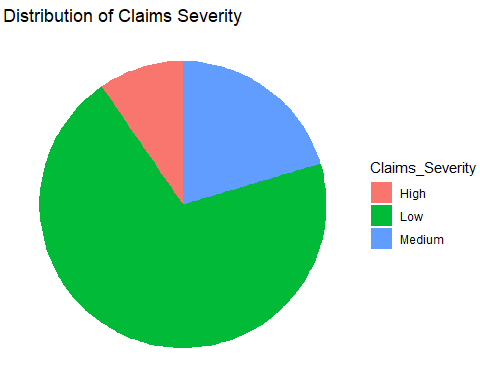
insurance\_data %>%  
 group\_by(Region) %>%  
 summarise(Count = n()) %>%  
 ggplot(aes(x = "", y = Count, fill = Region)) +  
 geom\_col(width = 1) +  
 coord\_polar(theta = "y") +  
 labs(title = "Distribution of Customers Across Regions") +  
 theme\_void()



*We displayed the customer distribution across different regions using a pie chart to understand regional variations.*

#Claims Severity Distribution

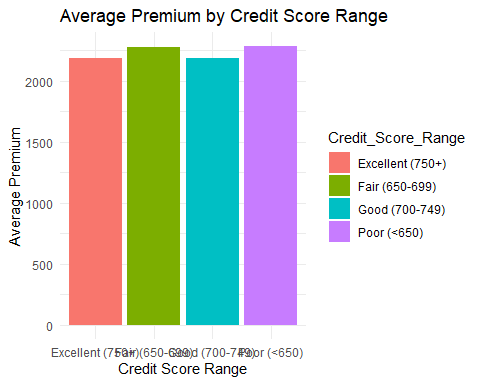
insurance\_data %>%  
 group\_by(Claims\_Severity) %>%  
 summarise(Count = n()) %>%  
 ggplot(aes(x = "", y = Count, fill = Claims\_Severity)) +  
 geom\_col(width = 1) +  
 coord\_polar(theta = "y") +  
 labs(title = "Distribution of Claims Severity") +  
 theme\_void()



*We visualized the distribution of claims severity levels using both a pie chart and a bar chart.*

#Premium Amount by Credit Score Range

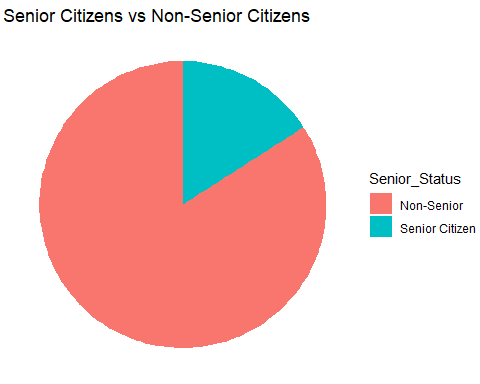
insurance\_data %>%  
 group\_by(Credit\_Score\_Range) %>%  
 summarise(Average\_Premium = mean(Premium\_Amount, na.rm = TRUE)) %>%  
 ggplot(aes(x = Credit\_Score\_Range, y = Average\_Premium, fill = Credit\_Score\_Range)) +  
 geom\_col() +  
 labs(title = "Average Premium by Credit Score Range", x = "Credit Score Range", y = "Average Premium") +  
 theme\_minimal()



*We showed how average premium amounts vary across different credit score ranges with a bar chart.*

#Senior vs Non-Senior Citizens Distribution

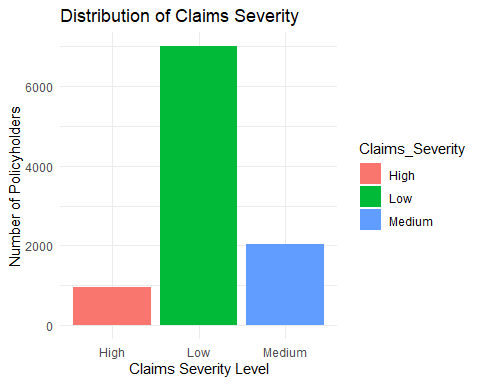
insurance\_data %>%  
 mutate(Senior\_Status = ifelse(Is\_Senior == 1, "Senior Citizen", "Non-Senior")) %>%  
 group\_by(Senior\_Status) %>%  
 summarise(Count = n()) %>%  
 ggplot(aes(x = "", y = Count, fill = Senior\_Status)) +  
 geom\_col(width = 1) +  
 coord\_polar(theta = "y") +  
 labs(title = "Senior Citizens vs Non-Senior Citizens") +  
 theme\_void()



*We compared the number of senior and non-senior citizens using a pie chart to understand the senior citizen customer base.*

#Claims Severity Distribution

insurance\_data %>%  
 group\_by(Claims\_Severity) %>%  
 summarise(Count = n()) %>%  
 ggplot(aes(x = Claims\_Severity, y = Count, fill = Claims\_Severity)) +  
 geom\_col() +  
 labs(title = "Distribution of Claims Severity", x = "Claims Severity Level", y = "Number of Policyholders") +  
 theme\_minimal()

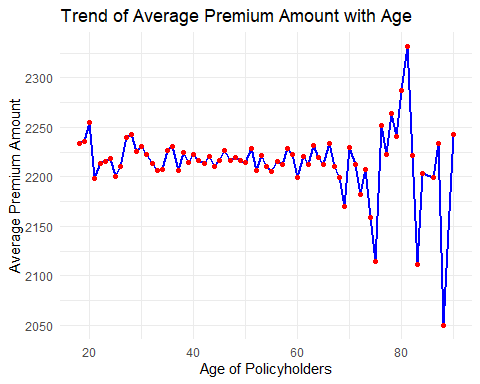


*We visualized claims severity distribution again through a bar chart for a different perspective.*

#How does the Average Premium Amount change with increasing Age?

age\_premium\_trend <- insurance\_data %>%  
 group\_by(Age) %>%  
 summarise(Average\_Premium = mean(Premium\_Amount, na.rm = TRUE))  
  
ggplot(age\_premium\_trend, aes(x = Age, y = Average\_Premium)) +  
 geom\_line(color = "blue", size = 1) +  
 geom\_point(color = "red", size = 1.5) +  
 labs(title = "Trend of Average Premium Amount with Age",  
 x = "Age of Policyholders",  
 y = "Average Premium Amount") +  
 theme\_minimal()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

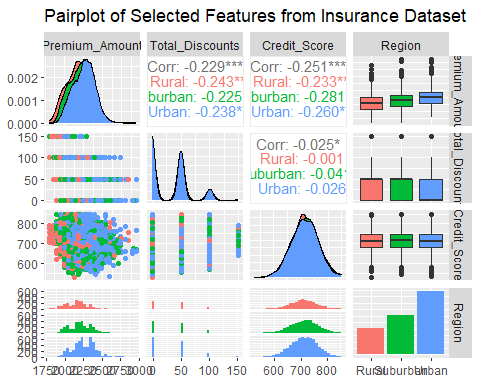


*The line chart shows how the average premium amount changes as the age of policyholders increases.* *A rising trend would suggest that older customers generally pay higher premiums, possibly due to increased insurance risks.* *Any dips might suggest discounts or different premium structures for certain age groups.*

# Pairplot of Selected Features

insurance\_subset <- insurance\_data[, c("Premium\_Amount", "Total\_Discounts", "Credit\_Score", "Region")]  
ggpairs(insurance\_subset,  
 columns = 1:4,  
 aes(color = Region),  
 title = "Pairplot of Selected Features from Insurance Dataset")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 *Finally, we created a pairplot of selected features — Premium Amount, Total Discounts, Credit Score, and Region — to study the relationships between these important variables.*