LOVELY PROFESSIONAL UNIVERSITY

Course: Data Exploration and Preparation

Course Code: CAP482

CA 3

Dated: - 28/April/2025

Submitted by

Name1: Anshu Kumar
Reg:12303955
Name2: Sonu Kumar
Reg:12303430
Section: DE225, Group: 1

Submitted to

Ms. Ranjit Kaur Walia
UID: 28632
Assistant Professor
SCA, LPU

Lovely Faculty of Technology & Sciences

School of Computer Applications

Lovely Professional University

Punjab

Synthetic Insurance Analysis

Anshu Kumar & Sonu Kumar

2025-04-27

Title: Analyzing Insurance Data Using R

Project Overview

This project analyzes insurance data to study customer profiles, premium distributions, claims behavior, and the impact of demographic and financial factors on insurance outcomes.
The aim is to uncover patterns related to age, credit score, marital status, prior insurance, and claims adjustment to help understand insurance risk better.

Dataset Used

The dataset contains information about policyholders: Age, Marital Status, Region, Premium Amount, Claims Frequency, Claims Severity, Credit Score, Prior Insurance, and Website Visits.

Rpubs Link: https://rpubs.com/anshh_k/insurance-analysis

Objectives

1. Understand the dataset (columns, data types, missing values).
2. Perform basic statistical analysis.
3. Explore correlations among important variables.
4. Identify patterns across demographic and financial features.
5. Visualize key insights using charts and plots.

#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 (H_0): There is no difference in Premium Amount between marital status groups. Alternative Hypothesis (H_1): 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 (H_0): The mean Claims Frequency is the same for all Policy Types. Alternative Hypothesis (H_1): 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

```r
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
```

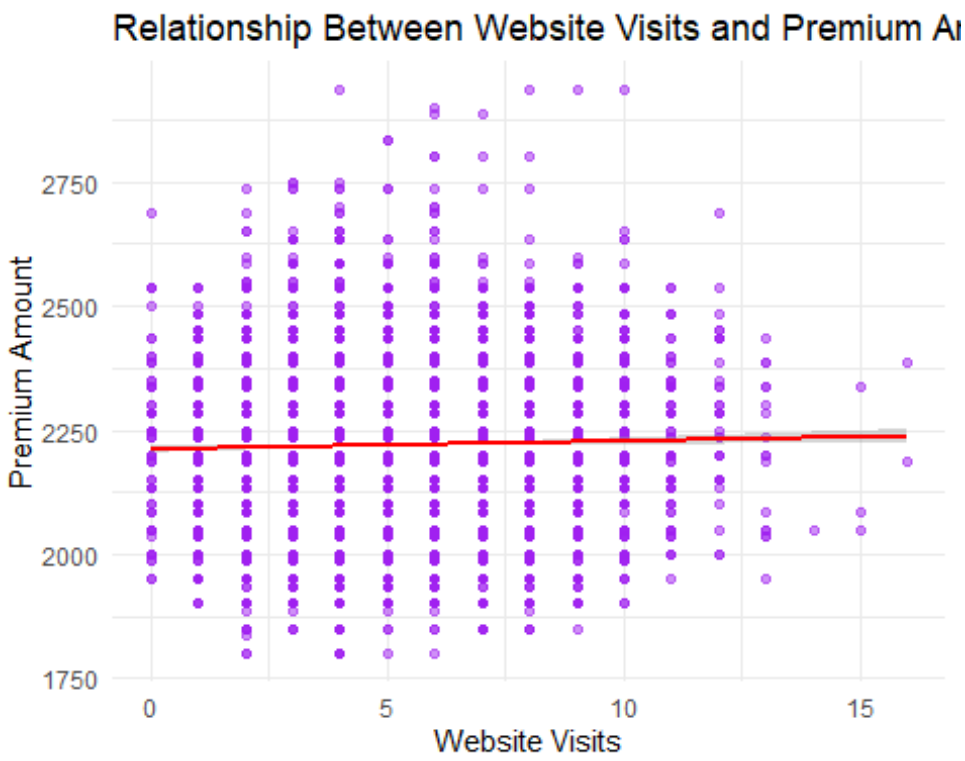
We performed a one-way ANOVA to investigate whether the Premium Amount significantly differs across different Regions. Null Hypothesis (H_0): The mean Premium Amount is the same across all regions. Alternative Hypothesis (H_1): At least one region has a different mean Premium Amount. A p-value less than 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."

6. Data Visualization

##Website Visits vs Premium Amount

```r
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()
```

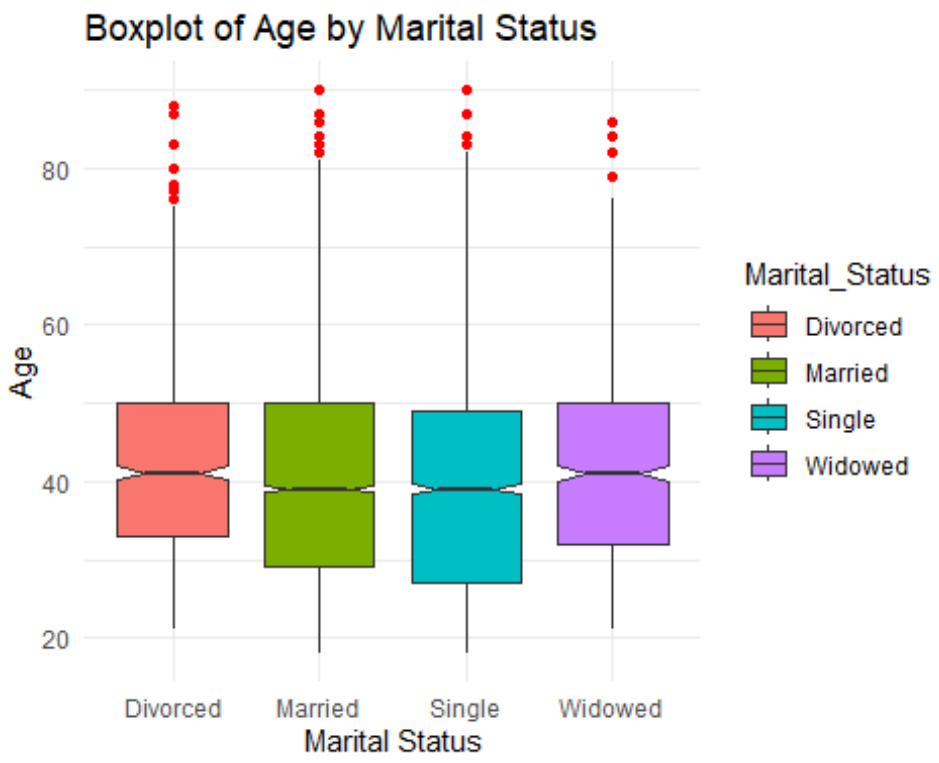
```
## `geom_smooth()` using formula = 'y ~ x'
```



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

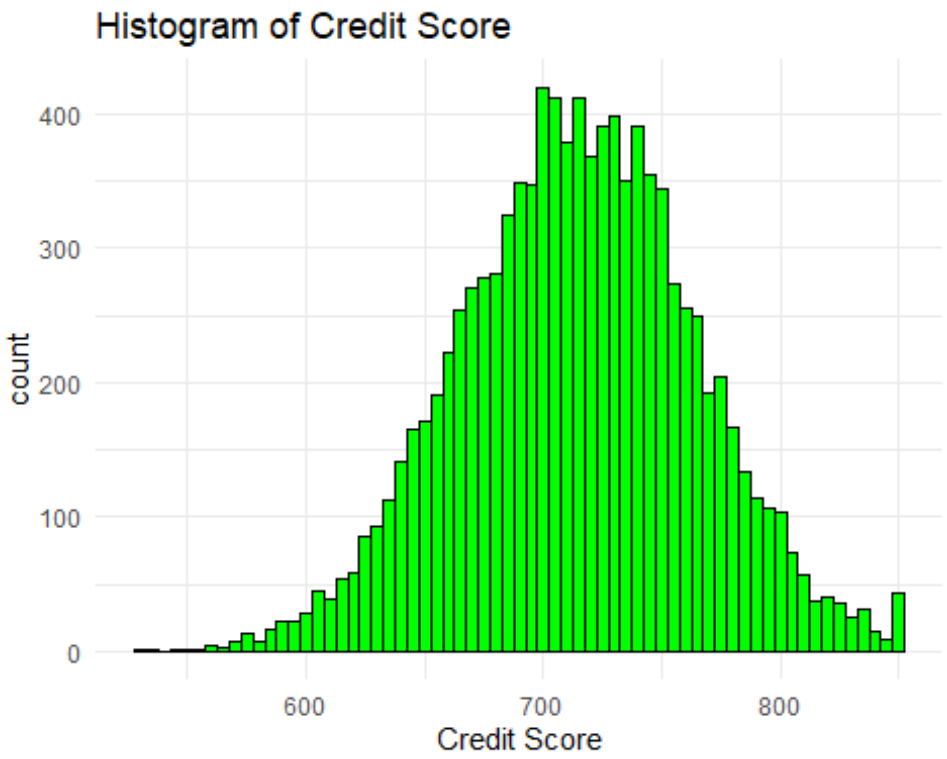
```
ggplot(insurance_data, aes(x = Marital_Status, y = Age, fill =
Marital_Status)) +
  geom_boxplot(outlier.color = "red", notch = TRUE) +
  labs(title = "Boxplot of Age by Marital Status", x = "Marital Status", y =
"Age") +
  theme_minimal()
```



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

Histogram of Credit Score

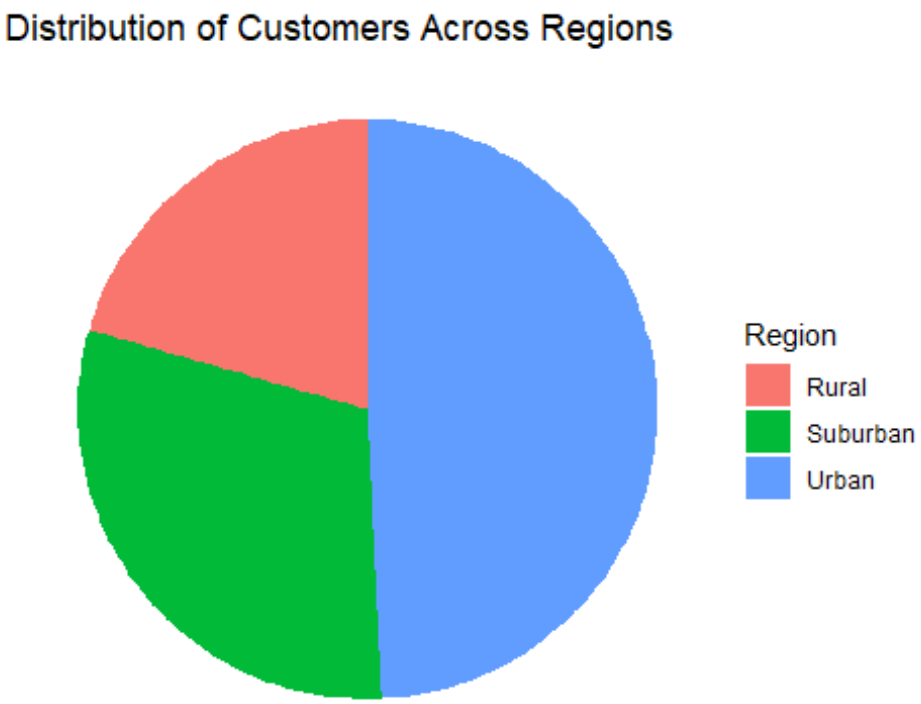
```
ggplot(insurance_data, aes(x = Credit_Score)) +
  geom_histogram(binwidth = 5, fill = "green", color = "black") +
  labs(title = "Histogram of Credit Score", x = "Credit Score") +
  theme_minimal()
```



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

#Customer Distribution across Regions

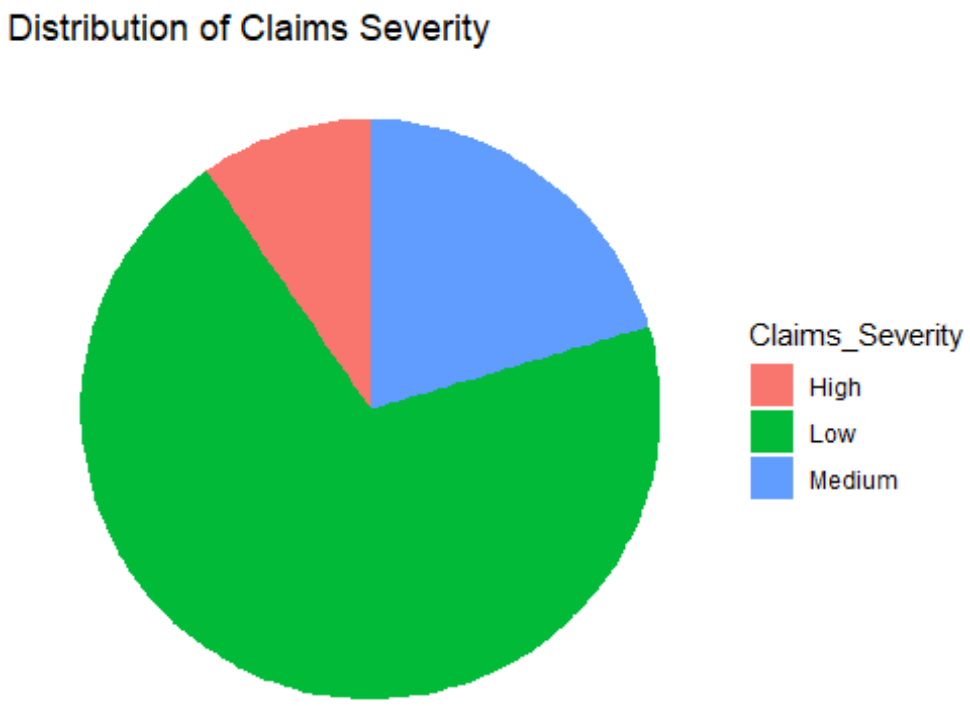
```
insurance_data %>%
  group_by(Region) %>%
  summarise(Count = n()) %>%
  ggplot(aes(x = "", y = Count, fill = Region)) +
  geom_col(width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Distribution of Customers Across Regions") +
  theme_void()
```



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

#Claims Severity Distribution

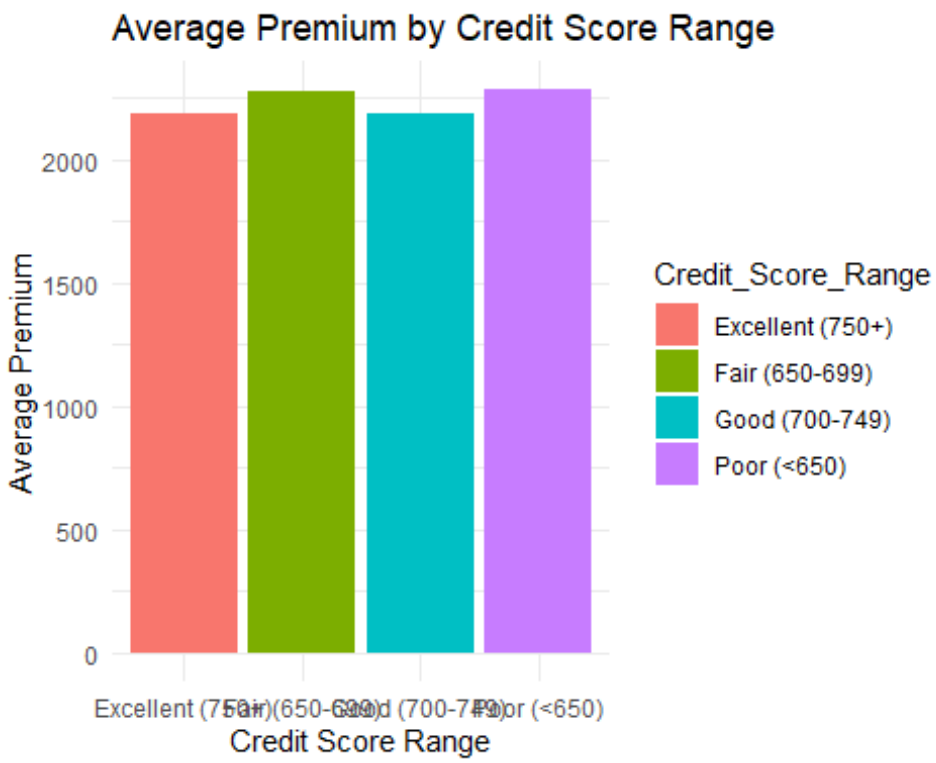
```
insurance_data %>%
  group_by(Claims_Severity) %>%
  summarise(Count = n()) %>%
  ggplot(aes(x = "", y = Count, fill = Claims_Severity)) +
  geom_col(width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Distribution of Claims Severity") +
  theme_void()
```



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

#Premium Amount by Credit Score Range

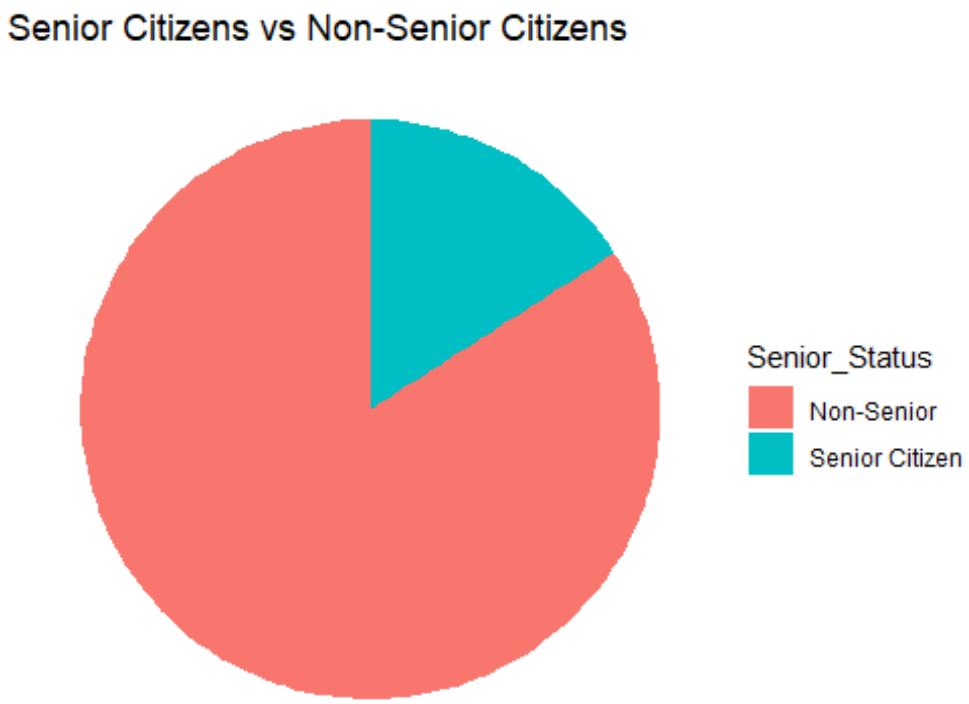
```
insurance_data %>%
  group_by(Credit_Score_Range) %>%
  summarise(Average_Premium = mean(Premium_Amount, na.rm = TRUE)) %>%
  ggplot(aes(x = Credit_Score_Range, y = Average_Premium, fill =
Credit_Score_Range)) +
  geom_col() +
  labs(title = "Average Premium by Credit Score Range", x = "Credit Score
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

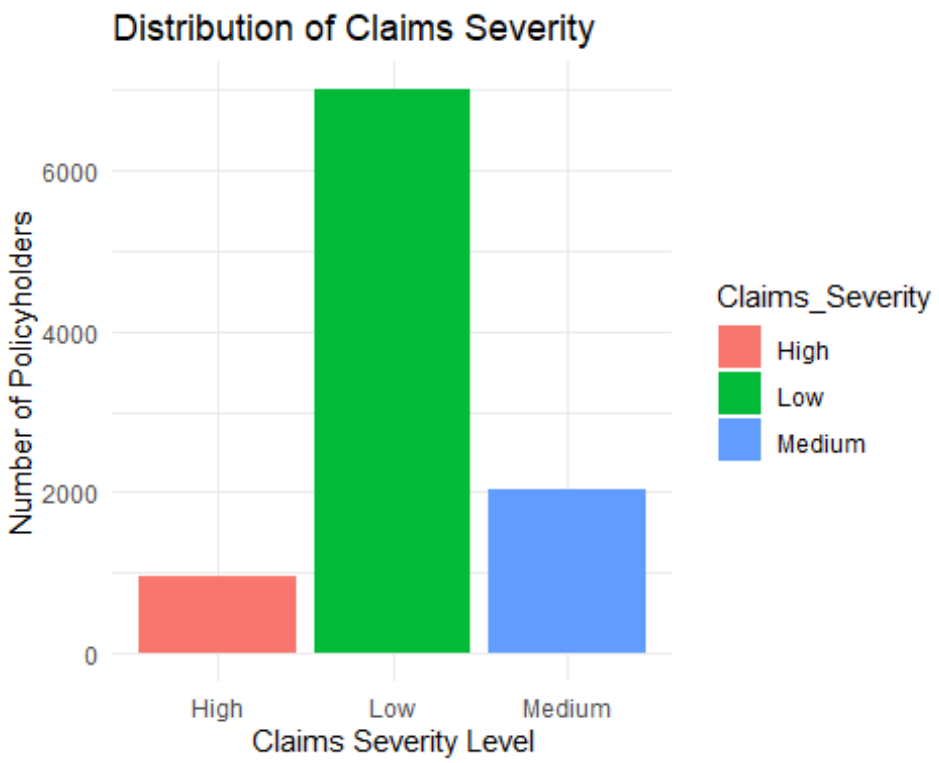
```
insurance_data %>%
  mutate(Senior_Status = ifelse(Is_Senior == 1, "Senior Citizen",
"Non-Senior")) %>%
  group_by(Senior_Status) %>%
  summarise(Count = n()) %>%
  ggplot(aes(x = "", y = Count, fill = Senior_Status)) +
  geom_col(width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Senior Citizens 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()
```



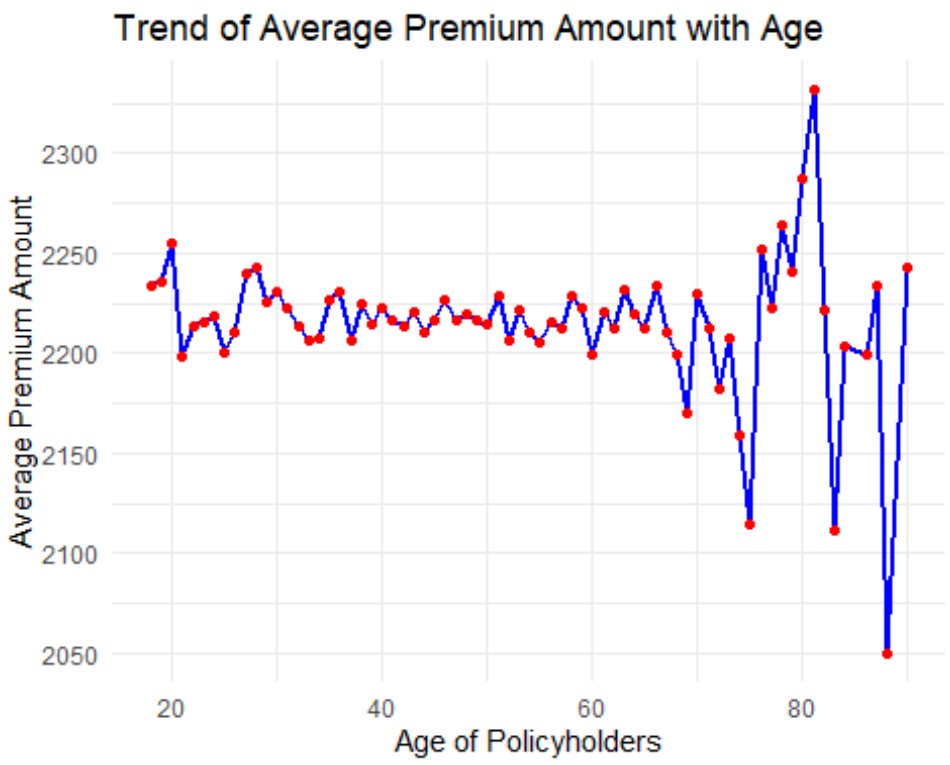
We visualized claims severity distribution again through a bar chart for a different perspective.

#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.
```

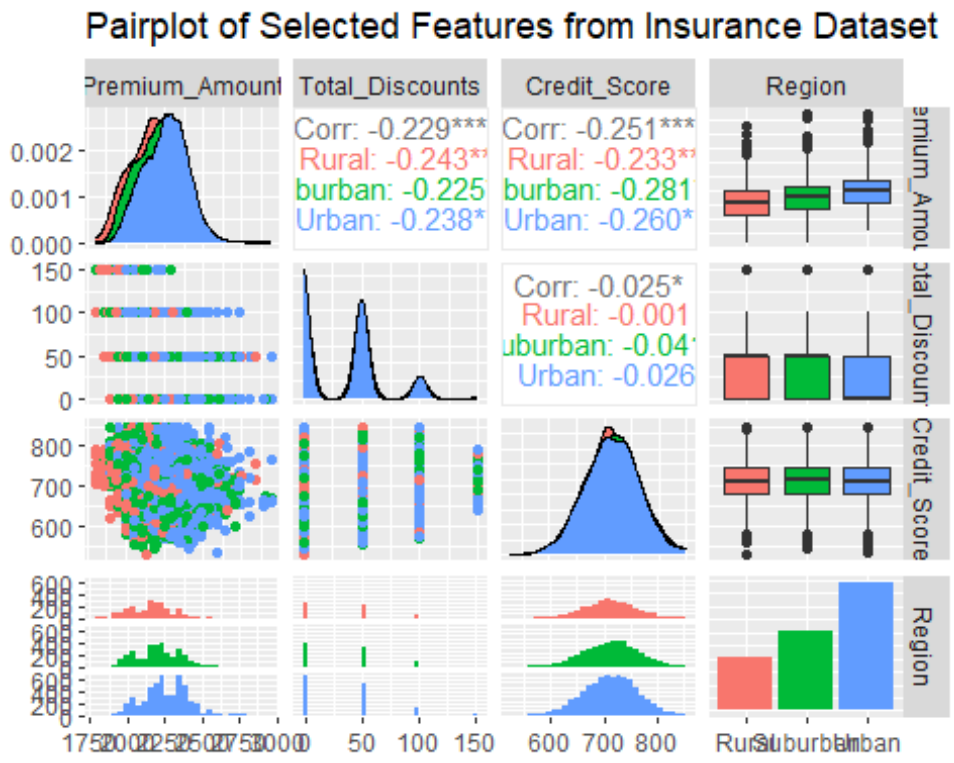


The line chart shows how the average premium amount changes as the age of policyholders increases. A rising trend would suggest that older customers generally pay higher premiums, possibly due to increased insurance risks. Any dips might suggest discounts or different premium structures for certain age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