Final Project Report
Dataset: Health Insurance
Neha Save

HEALTH INSURANCE DATA ANALYSIS

Introduction

The Health Insurance Dataset Analysis project is an in-depth study that looks at the different factors affecting how much people claim on their health insurance. This project uses a large set of data that includes information like a person's age, whether they're male or female, their weight, Body Mass Index (BMI), if they have any inherited diseases, whether they smoke or not, how often they exercise, and other details. These pieces of information help us understand what influences the cost of health insurance claims.

The main goal of this project is to figure out how different health-related factors, such as how old you are, your lifestyle, and if you have certain health conditions, impact the amount of money claimed on health insurance. This work is important because it can help insurance companies, people who buy insurance, and doctors understand how different things about your health and life can make your insurance claims go up or down.

To do this analysis, the project uses several advanced math techniques to look at how each piece of information (like age or if you smoke) affects insurance claims. It pays special attention to things like whether someone smokes or exercises regularly, and how these choices relate to the cost of claims. It also looks into how being older or having a higher BMI can increase the risk of getting diabetes, which is another factor that can affect insurance costs.

The project also deals with challenges like missing information or very high or low numbers that don't fit the pattern (called outliers), to make sure the results are as accurate as possible.

By examining all these factors, this study helps us see clearly which behaviors and health conditions are most likely to lead to higher insurance claims. This is useful for everyone involved, from the people who buy insurance to the ones who provide it, because it shows how making healthier choices can potentially lower the cost of insurance claims.

In simple terms, this project helps us understand the connection between our health, our choices, and how much we might need to claim on our health insurance. It's a way of using a lot of data to find out more about how our health and lifestyle choices can affect insurance costs, which can help make insurance fairer and encourage us to live healthier lives.

Dataset Overview

The Health Insurance Dataset comprises several key variables that together provide a comprehensive view of factors influencing insurance claims. Below is an overview of the dataset's columns, each representing a unique attribute of the individuals:

```
> # Head of the dataset
> head(insurance_data)
# A tibble: 6 x 13
    age sex    weight   bmi hereditary_diseases no_of_dependents smoker city        bloodpressure diabetes
  <dbl> <chr>   <dbl> <dbl> <chr>                          <dbl>  <dbl> <chr>               <dbl>    <dbl>
1    60 male       64  24.3 NoDisease                          1      0 NewYork                72        0
2    49 female     75  22.6 NoDisease                          1      0 Boston                 78        1
3    32 female     64  17.8 Epilepsy                           2      1 Phildelphia            88        1
4    61 female     53  36.4 NoDisease                          1      1 Pittsburg              72        1
5    19 female     50  20.6 NoDisease                          0      0 Buffalo                82        1
6    42 female     89  37.9 NoDisease                          0      0 AtlanticCity           78        0
# i 3 more variables: regular_exercise <dbl>, job_title <chr>, claim <dbl>
```

Age: Reflects individual's age, crucial for assessing health risk and insurance cost.

Sex: Gender of individuals, affecting risk assessment and health needs.

Weight: Individual's weight, indicative of overall health status.

BMI: Measures obesity levels, impacts health risk and insurance premiums.

Hereditary Diseases: Presence of inherited diseases affects insurance risk assessments.

No. of Dependents: Influences insurance coverage needs and potential claims.

Smoker: Smoking status significantly impacts health risks and insurance rates.

City: Residence city affects healthcare access and cost variations.

Blood Pressure: Indicator of cardiovascular health affects risk and premiums.

Diabetes: Presence of diabetes, critical for assessing health and coverage.

Regular Exercise: Exercise habits, impacts health status and insurance costs.

Job Title: Provides insight into occupational risks and lifestyle.

Claim: Insurance claim amount reflects financial implications of health risks.

Analysis

Summary:

```
> summary(insurance_data)
      age            sex                weight           bmi        hereditary_diseases no_of_dependents
 Min.   :18.00   Length:15000       Min.   :34.00   Min.   :16.00   Length:15000        Min.   :0.00
 1st Qu.:27.00   Class :character   1st Qu.:54.00   1st Qu.:25.90   Class :character    1st Qu.:0.00
 Median :40.00   Mode  :character   Median :63.00   Median :29.80   Mode  :character    Median :1.00
 Mean   :39.55                      Mean   :64.91   Mean   :30.27                       Mean   :1.13
 3rd Qu.:51.00                      3rd Qu.:76.00   3rd Qu.:34.10                       3rd Qu.:2.00
 Max.   :64.00                      Max.   :95.00   Max.   :53.10                       Max.   :5.00
     smoker           city           bloodpressure       diabetes     regular_exercise  job_title
 Min.   :0.0000   Length:15000       Min.   :  0.00   Min.   :0.000   Min.   :0.0000   Length:15000
 1st Qu.:0.0000   Class :character   1st Qu.: 64.00   1st Qu.:1.000   1st Qu.:0.0000   Class :character
 Median :0.0000   Mode  :character   Median : 71.00   Median :1.000   Median :0.0000   Mode  :character
 Mean   :0.1981                     Mean   : 68.65   Mean   :0.777   Mean   :0.2241
 3rd Qu.:0.0000                     3rd Qu.: 80.00   3rd Qu.:1.000   3rd Qu.:0.0000
 Max.   :1.0000                     Max.   :122.00   Max.   :1.000   Max.   :1.0000
     claim
 Min.   : 1122
 1st Qu.: 4847
 Median : 9546
 Mean   :13401
 3rd Qu.:16519
 Max.   :63770
```

The summary gives us a good overall picture of our dataset. It's like a snapshot showing us the average values, the middle values, and the range of our key features. We noticed that there are some missing values in 'age' and 'bmi', which means we might need to fill in those gaps or take a closer look at those entries. Looking at things like age, weight, BMI, and claim amounts helps us understand how our data is spread out. We also have some categorical info about 'sex', 'hereditary_diseases', 'city', and 'job_title', giving us a glimpse into the personal and lifestyle aspects of our dataset. The 'claim' variable shows a wide range of amounts, hinting that there's a lot of variation we might want to explore in future analysis.

Missing Values:

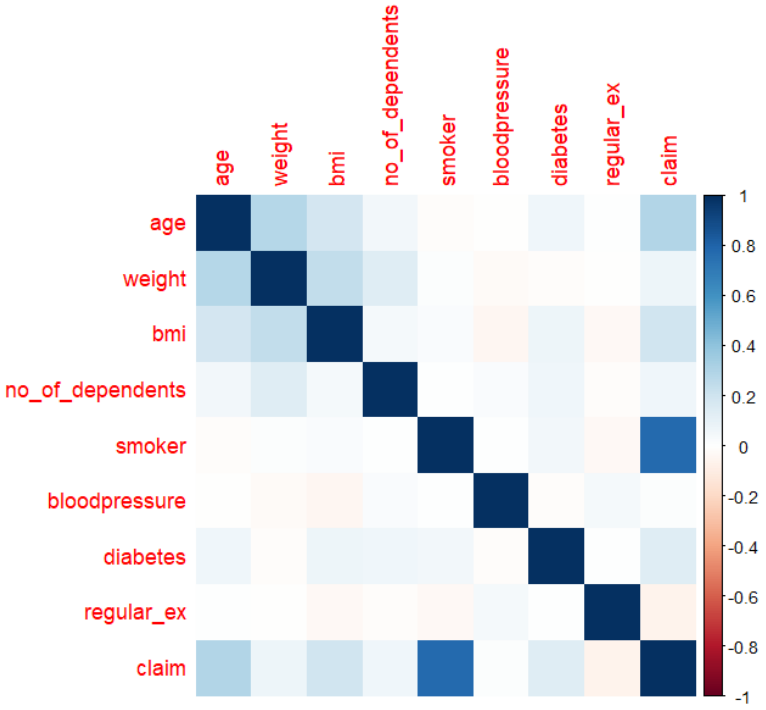
```
> # Check for missing values in each column
> missing_values_column <- colSums(is.na(insurance_data))
> print(missing_values_column)
               age                sex             weight                bmi hereditary_diseases
               396                  0                  0                956                  0
  no_of_dependents             smoker               city       bloodpressure            diabetes
                 0                  0                  0                  0                  0
        regular_ex          job_title              claim
                 0                  0                  0
> # Impute missing values for 'age' with mean
> insurance_data$age <- ifelse(is.na(insurance_data$age), mean(insurance_data$age, na.rm = TRUE), insurance_data
$age)
>
> # Impute missing values for 'bmi' with mean
> insurance_data$bmi <- ifelse(is.na(insurance_data$bmi), mean(insurance_data$bmi, na.rm = TRUE), insurance_data
$bmi)
> # Check for missing values
> sum(is.na(insurance_data))
[1] 0
```

In the process of preparing our dataset for analysis, we encountered missing values in the 'age' and 'bmi' columns. To handle this, we decided to impute these missing values with the mean values of their respective columns, ensuring that our dataset remains

comprehensive and representative. Through these measures, we addressed data inconsistencies and made our dataset more robust for subsequent analysis.

Correlation Matrix:

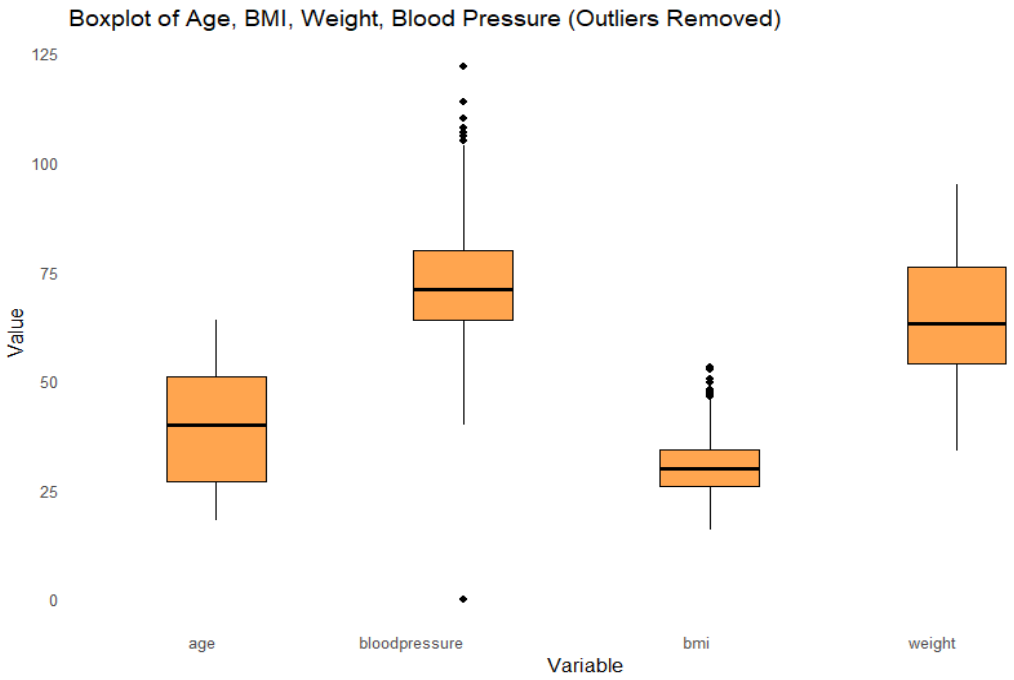
```
> # Correlation matrix for numeric variables
> correlation_matrix <- cor(insurance_data %>% select_if(is.numeric))
> print(correlation_matrix)
                      age       weight         bmi no_of_dependents      smoker bloodpressure
age            1.000000000  0.281227763  0.18012275       0.05892611 -0.01941711  -0.008225211
weight         0.281227763  1.000000000  0.24326927       0.13568750  0.01549894  -0.020835013
bmi            0.180122750  0.243269268  1.00000000       0.04970435  0.02280168  -0.040065742
no_of_dependents 0.058926106 0.135687497 0.04970435       1.00000000  0.00836426   0.024848503
smoker        -0.019417105  0.015498937  0.02280168       0.00836426  1.00000000   0.005709030
bloodpressure -0.008225211 -0.020835013 -0.04006574       0.02484850  0.00570903   1.000000000
diabetes       0.069597515 -0.010489593  0.07906611       0.06518228  0.05816377  -0.016498340
regular_ex     0.008447583 -0.005578157 -0.03791996      -0.01030237 -0.03694930   0.042492812
claim          0.298359470  0.077716218  0.19793925       0.06761373  0.77339865   0.013742209
                  diabetes   regular_ex       claim
age            0.069597515  0.008447583  0.29835947
weight        -0.010489593 -0.005578157  0.07771622
bmi            0.079066108 -0.037919955  0.19793925
no_of_dependents 0.065182278 -0.010302366 0.06761373
smoker         0.058163766 -0.036949295  0.77339865
bloodpressure -0.016498340  0.042492812  0.01374221
diabetes       1.000000000  0.007960027  0.13537119
regular_ex     0.007960027  1.000000000 -0.06049183
claim          0.135371186 -0.060491827  1.00000000
```



The correlation matrix offers a fascinating glimpse into the connections among different aspects of our health insurance dataset. Notably, age shows a positive connection with weight (0.28) and BMI (0.18), indicating that as age increases, weight and BMI tend to follow suit. There's also a positive link between weight and BMI (0.24), suggesting that individuals with higher weights tend to have higher BMI values. On a different note, blood pressure seems to maintain a relatively weak correlation with age, weight, and BMI. As we

go deeper, we find that regular exercise demonstrates a subtle negative correlation with BMI (-0.04) and a more noticeable negative link with insurance claim amounts (-0.06). However, the most striking revelation lies in the relationship between smoking and insurance claims, boasting a substantial correlation of 0.77. This implies that smokers tend to have notably higher health insurance claims.

Addressing Outliers in the Dataset:



Upon examining the initial boxplot for age, BMI, weight, and blood pressure in the insurance dataset, it was observed that there were outliers present, particularly in the 'bmi' and 'bloodpressure' variables. To address this, we proceeded to remove these outliers using the Interquartile Range (IQR) method.

Following the removal of outliers, the cleaned data was reshaped for visualization. The resulting boxplot of age, BMI, Weight, Blood Pressure depicts the distribution of these variables without the influence of outliers. Upon inspection of the new boxplot, it became apparent that outliers in both 'bmi' and 'bloodpressure' were successfully eliminated.

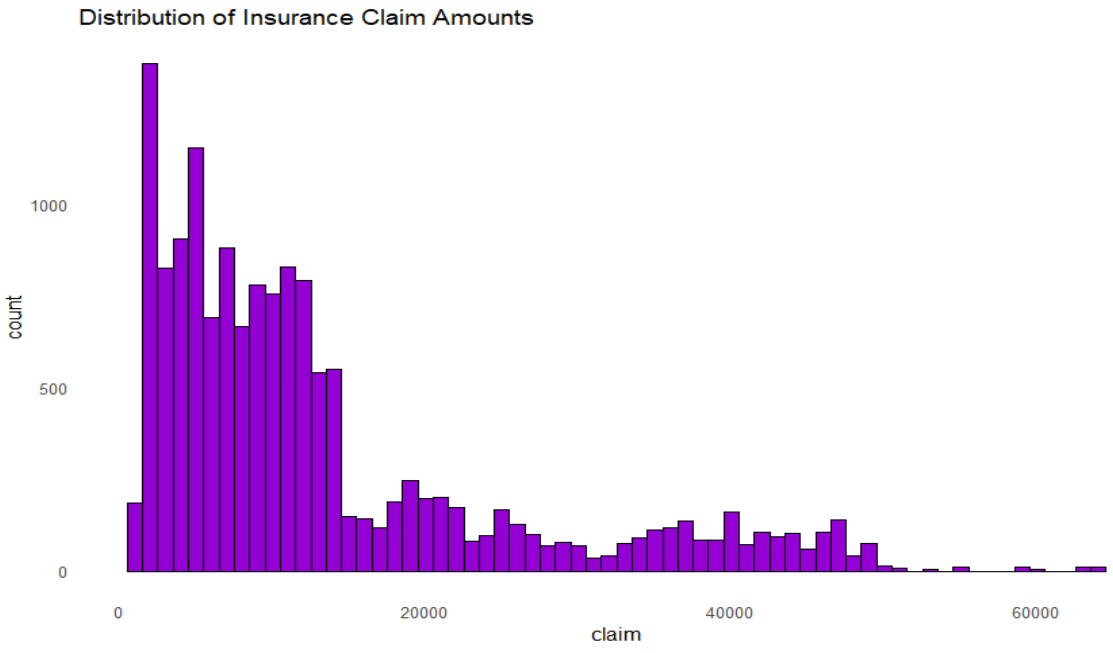
Transforming Categorical Data:

```
>   # Convert 'sex' to binary (1 for 'male', 0 for 'female')
>   final_dataset$sex <- as.numeric(final_dataset$sex  == "male")
> View(final_dataset)
>  final_dataset$sex
   [1] 1 0 0 0 0 0 1 1 1 0 0 0 1 0 1 0 0 1 1 0 1 0 0 1 0 0 1 0 0 1 1 0 1 1 0 1 1 0 1 0 0 0 1 1 0 0 1 0 1 0 0
  [52] 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 0 0 0 0 0 1 1 0 0 1 1 1 0 1 0 0 1 1 0 1 0 1 1 0 0 1 0 0 0 0 0 0 1 0 0
 [103] 0 1 1 1 0 1 0 1 0 0 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 1 0
 [154] 0 1 1 1 0 0 0 0 0 1 1 1 0 0 1 0 1 1 0 1 0 0 1 0 1 1 0 0 0 1 1 1 0 1 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0
 [205] 1 1 1 1 0 0 1 0 1 1 0 0 1 1 0 1 0 1 1 1 0 1 1 1 0 0 1 1 0 0 0 0 0 0 0 0 1 1 0 1 1 1 0 1 0 0 1 0 0 1 0
 [256] 0 1 0 0 1 0 1 0 1 0 0 1 0 1 1 1 0 0 0 0 0 0 0 1 1 0 1 1 0 1 1 0 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0 0 0 1 0 0
 [307] 0 0 0 0 1 1 0 1 0 1 0 0 0 1 1 0 0 1 1 1 0 0 0 0 1 1 0 1 1 0 1 0 0 1 0 0 1 1 0 1 1 1 1 0 1 1 1 1 1 0 0
 [358] 1 0 1 1 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 1 0 0 0 1 1 0 0 1 0 1 1 1 1 1 1 0 0 1 0 1 1 1 1 0 0 0 1 1 0 1 0 0
 [409] 1 0 1 0 1 1 1 1 0 1 0 0 0 0 1 1 0 1 0 1 1 1 1 0 0 1 0 1 1 0 0 0 0 1 1 0 1 1 1 1 1 0 1 1 1 0 0 1 1 0 0
 [460] 0 0 1 1 0 0 1 0 0 1 0 1 1 1 1 1 0 1 0 0 0 0 0 0 0 0 1 0 1 1 0 0 1 1 0 1 0 0 1 0 0 1 0 0 0 1 1 1 0 1 0 1
 [511] 0 1 1 1 0 1 1 0 0 0 1 1 0 0 1 1 1 1 0 0 0 0 1 1 0 1 0 0 1 1 0 0 0 1 1 1 1 1 0 0 0 1 1 1 0 1 0 0 0 0 0
 [562] 0 1 1 0 1 1 0 1 0 0 1 1 0 1 0 1 1 0 1 0 0 0 1 0 1 1 1 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 1 0 1 1 0 1 0 0 1
 [613] 0 0 1 1 0 1 0 1 1 1 0 0 0 0 0 1 1 0 1 1 0 0 1 1 0 1 1 0 0 1 0 0 1 1 1 1 0 1 0 1 0 1 1 0 0 1 0 1 0 1 0
 [664] 1 1 1 0 0 0 0 0 1 1 0 1 1 0 0 0 0 0 1 0 1 0 1 0 0 1 0 1 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0
 [715] 0 0 0 0 0 1 1 0 0 1 0 1 0 0 0 1 1 0 1 1 1 1 1 1 1 0 1 0 0 1 0 1 1 1 1 0 1 0 0 0 1 0 1 1 1 1 0 0 0 0 0 1
 [766] 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 1 1 0 1 0 1 1 0 0 1 1 0 1 1 0 0 0 0 1 1 0 0 1 1 1 1 0 1 1 0 1 0 1 1 1 1
 [817] 0 1 1 0 0 1 0 0 1 0 1 1 0 0 1 0 0 0 0 1 0 1 1 1 0 1 1 1 1 0 0 0 0 1 0 1 1 1 0 0 0 1 1 1 1 0 0 1 0 1 1
 [868] 1 1 0 1 1 1 0 0 1 1 1 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 1 0 1 0 1 0 0 0 0 1 1 1 1 0 0 1 0 1 0 1 0 0 0 0 1 1 0
 [919] 1 0 1 1 0 0 0 0 0 1 1 1 1 1 0 0 0 1 0 0 0 0 0 1 1 1 0 1 1 0 0 0 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1 1 0 0 0 1
 [970] 1 1 1 1 0 1 0 1 0 0 1 1 1 0 1 1 1 1 0 0 0 0 1 0 1 0 1 0 1 0 1
 [ reached getOption("max.print") -- omitted 12967 entries ]
```

The 'sex' column in the dataset is categorical with values like "male" and "female,".It assigns the value 1 if the 'sex' is "male" and 0 if it is "female." In the resulting transformation, the 'sex' column will now contain 1 for "male" and 0 for "female."

```
>   # Apply mapping to 'hereditary_diseases' column in 'final_dataset'
>   final_dataset <- final_dataset %>%
+     mutate(hereditary_diseases = recode(hereditary_diseases, !!!hereditary_diseases_mapping))
>   final_dataset$hereditary_diseases
   [1] 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [52] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0
 [103] 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 6 0 0 0 0 0 0 0 0 0 0 0 0 0 7 0 0
 [154] 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 8 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [205] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [256] 0 6 0 0 0 0 0 7 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 9 0 0 9 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [307] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5 0 0 0 3 0 0 3 0 0 0 0 0 0 3 5 0 1 0 0 0 0 0 0 0 0 0 0
 [358] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 7 0 0 0 0 0 0 0 7 0 0
 [409] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0
 [460] 0 0 0 0 0 0 0 0 9 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [511] 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 0 0 0 0 0 0
 [562] 0 0 0 0 0 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 8 0 0 0 0 0 0 0 0 0
 [613] 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 6 0 0 0 0 0 0 0 0 0 0 0
 [664] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 1 5 0 2 0 0 0 0 0 0 0 0 0 0 0 0
 [715] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 7 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5 5 6 0 0 0 0 0 0 0 0 0 0 0 0
 [766] 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 9 0 3 0 0 0 0 0 0
 [817] 0 0 0 0 0 0 2 0 0 0 9 0 0 0 0 0 0 3 0 0 0 0 0 0 5 0 0 7 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 4
 [868] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 0 0 0 0 0 0 0
 [919] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 9 9 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0
 [970] 9 0 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0 0 3 0 0 0 0 0
 [ reached getOption("max.print") -- omitted 12967 entries ]
```

The hereditary diseases column defines a relationship between specific disease names and corresponding numeric codes. This mapping is useful when we want to represent categorical data numerically, making it easier to analyze or use in various computations.
Each disease name, such as 'Epilepsy,' 'HeartDisease,' etc., is associated with a unique numeric code from 0 to 9. For instance, 'NoDisease' is mapped to 0, 'Epilepsy' to 1, 'EyeDisease' to 2, and so on.

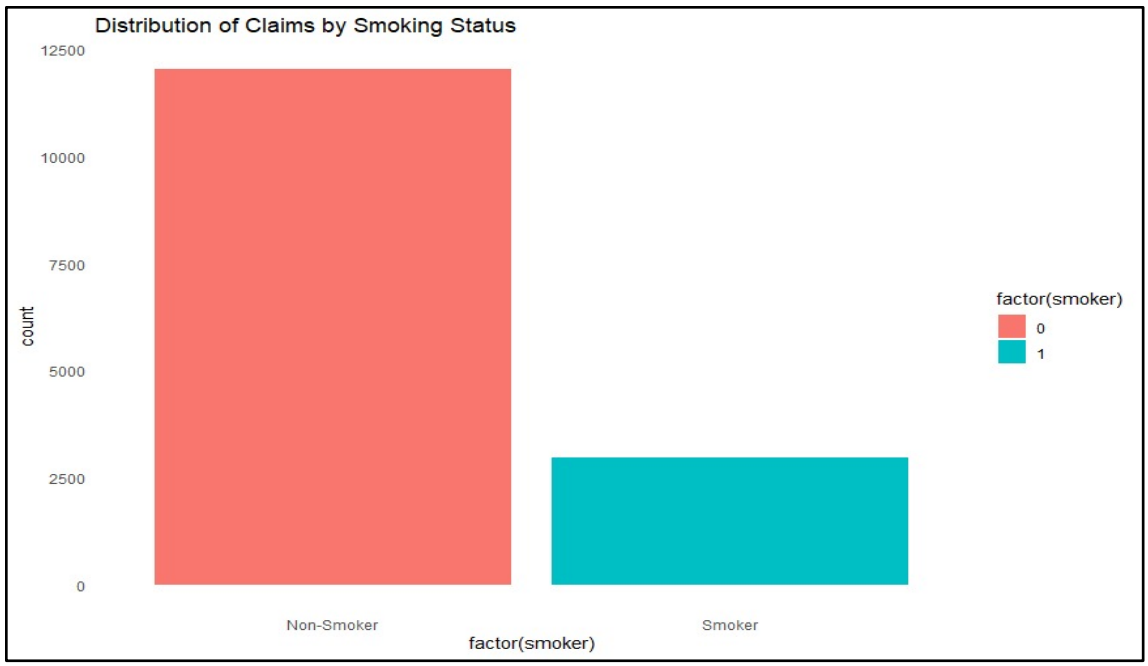


Histogram of Distribution of Claim Amounts

Upon examining the histogram, it is evident that most of the insurance claims fall within the range of 0 to 20,000 dollars. This range has the highest count, with more than 1000

occurrences, indicating a significant number of claims at lower amounts. As the claim amounts increase, the counts gradually decrease, forming a downward trend.

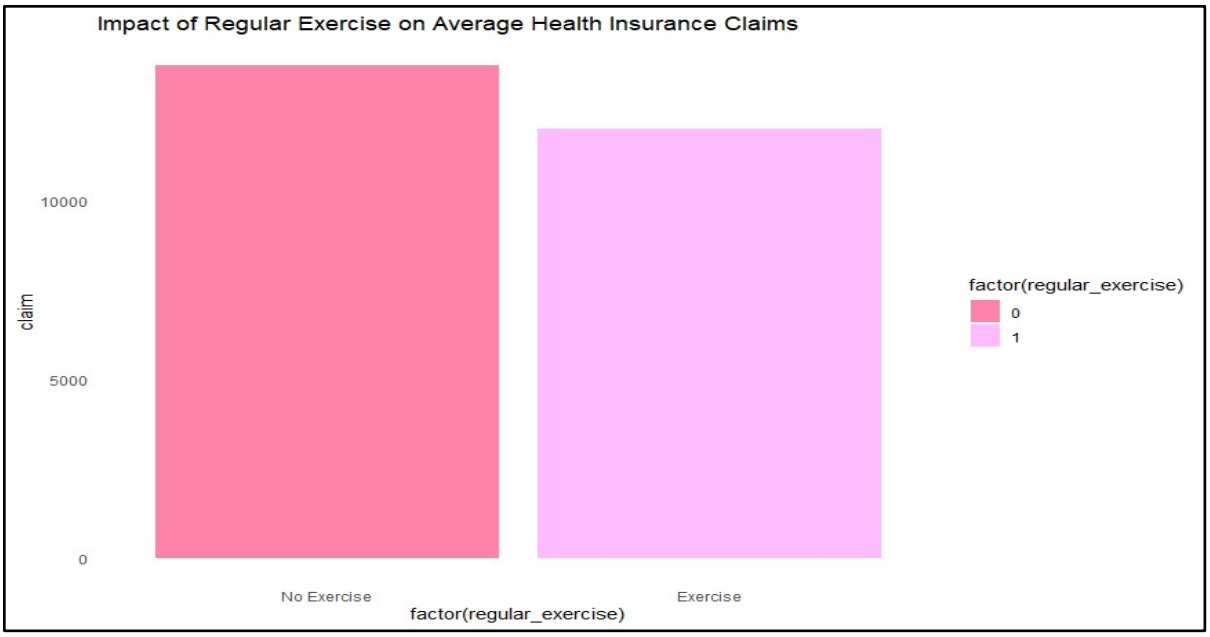
Specifically, focusing on the 20,000-to-50,000-dollar range, the histogram reveals a decline in the number of claims, with approximately 250 occurrences. This suggests the frequency of insurance claims decreases as the claim amounts rise within this range.



Bar Plot of Distribution of Claims by Smoking Status

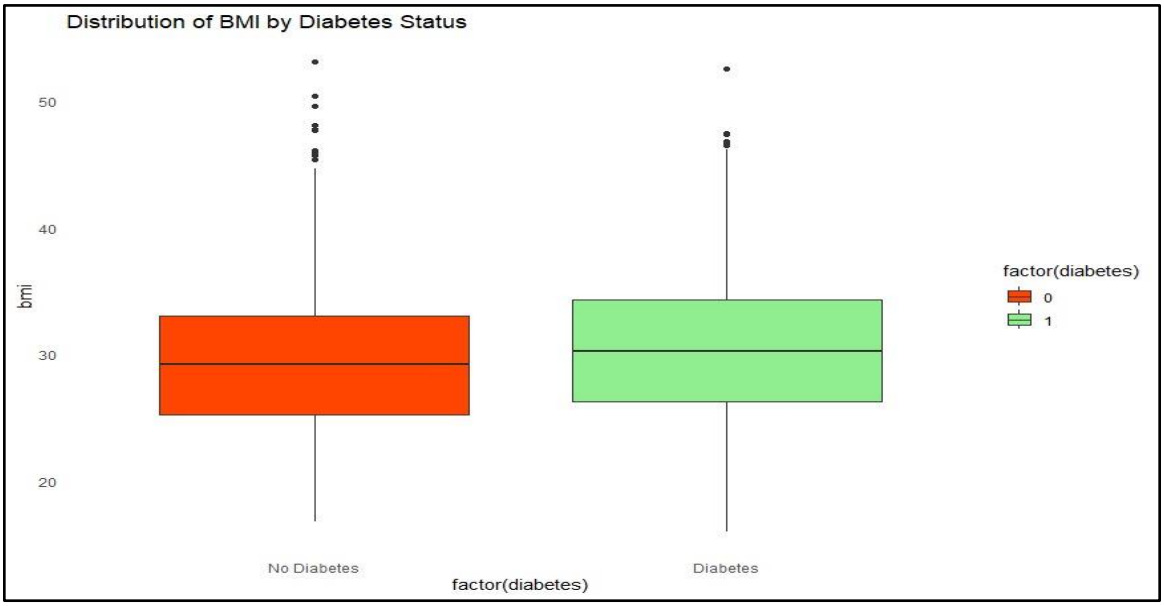
The interesting observation in the bar plot is that, for non-smokers, the bar is notably higher, suggesting a higher count of insurance claims in this group. Additionally, the associated claim amounts for non-smokers are particularly high, concentrated around 12,000. This implies that, despite the higher count of claims, the claim amounts for non-smokers tend to be substantial, indicating potentially more expensive or severe claims within this category.

On the other hand, for smokers, the bar is lower, suggesting a lower count of insurance claims in this group. The claim amounts for smokers are generally lower, centering around 3,000. This indicates that while the count of claims is lower for smokers, the associated claim amounts are also comparatively lower, suggesting a different distribution pattern in terms of both count and claim amounts.



Bar Plot of Imapact of Regular Exercise on Average Health Insurance Claims

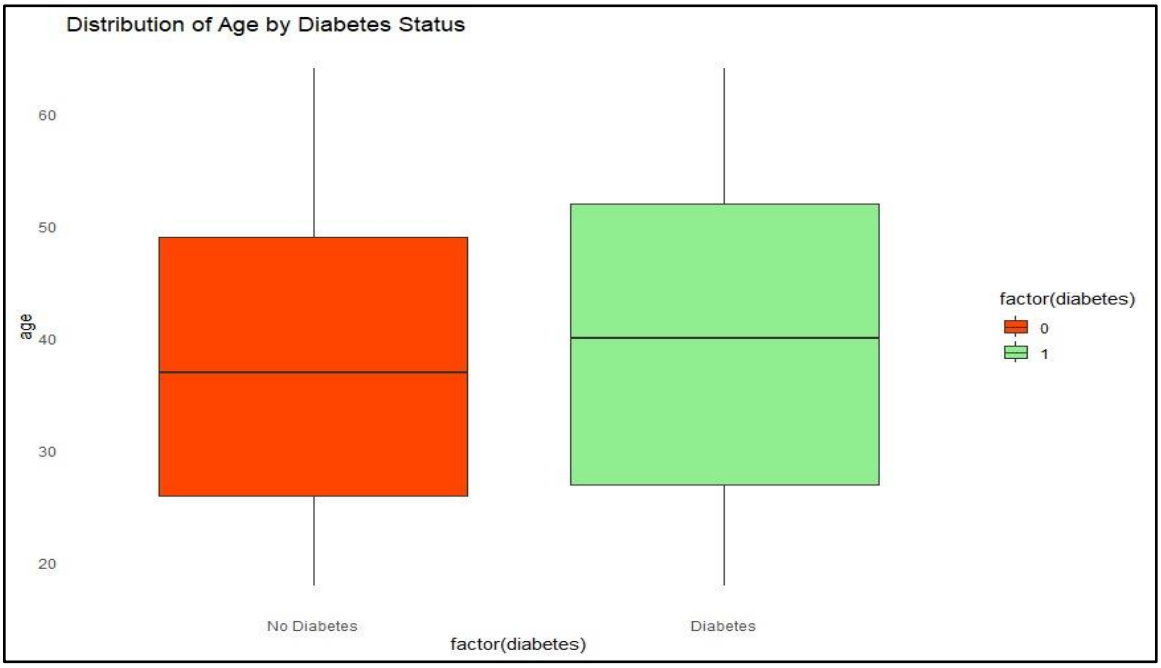
In this bar plot, a noticeable pattern emerges. For individuals without regular exercise ("No Exercise"), the bar is higher, indicating a higher average health insurance claim amount. This suggests that, on average, individuals who do not engage in regular exercise tend to have higher health insurance claims. Conversely, for individuals who engage in regular exercise ("Exercise"), the bar is lower, indicating a lower average health insurance claim amount. This implies that, on average, individuals with regular exercise habits tend to have lower health insurance claims.



Boxplot of BMI by Diabetes Status with Outliers

The boxplot visually illustrates the distribution of Body Mass Index (BMI) based on diabetes status, utilizing distinct colors for "No Diabetes" (0) and "Diabetes" (1) categories. For individuals without diabetes (0), the boxplot indicates that the upper quartile for BMI is approximately 33, the lower quartile is around 25, and the median BMI is roughly 29. This suggests that the BMI distribution for individuals without diabetes tends to be centered around 29, with a spread between 25 and 33.

Conversely, for individuals with diabetes (1), the boxplot shows that the upper quartile for BMI is about 35, the lower quartile is approximately 26, and the median BMI is around 30. This implies that the BMI distribution for individuals with diabetes is centered around 30, with a broader spread between 26 and 35.



Boxplot of BMI by Diabetes Status without Outliers

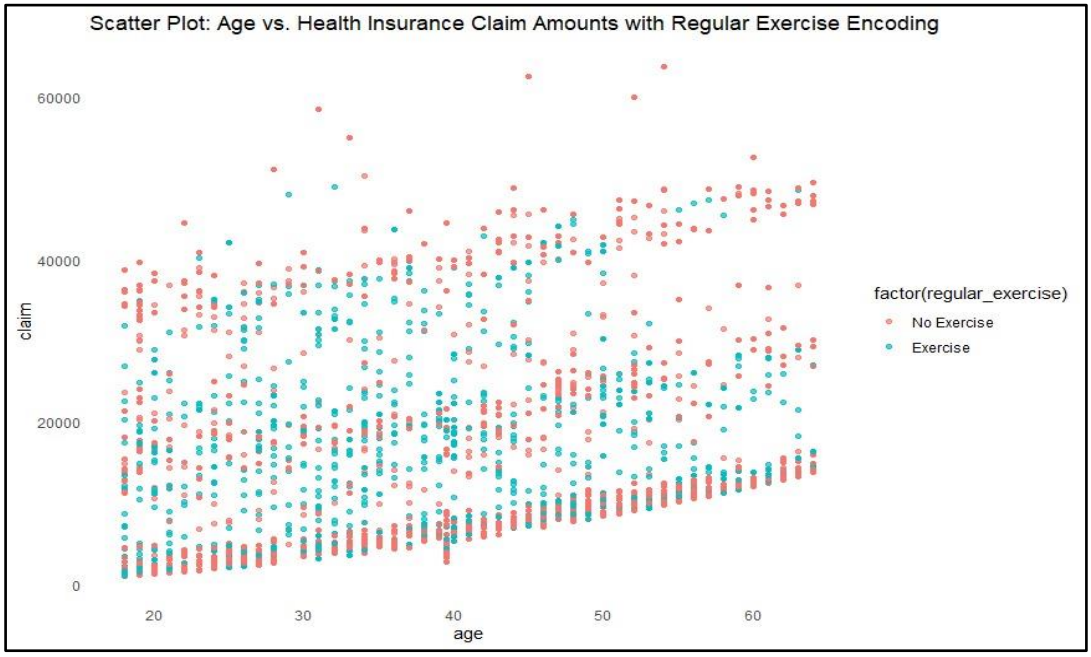
The boxplot visually represents the distribution of age based on diabetes status using different colors for "No Diabetes" and "Diabetes" categories. In "No Diabetes" (represented by 0), the upper quartile is about 50, the lower 25, and the median age is roughly 37. This suggests that the age distribution for individuals without diabetes tends to be centered around 37, with a spread between 25 and 50.

On the other hand, for individuals with diabetes (represented by 1), the boxplot indicates that the upper quartile is around 52, the lower quartile is approximately 26, and the median age is about 40. This implies that the age distribution for individuals with diabetes is centered around 40, with a wider spread between 26 and 52.

Questions

1. **How does age compare to other factors like weight or exercise in predicting health insurance claim amounts?**

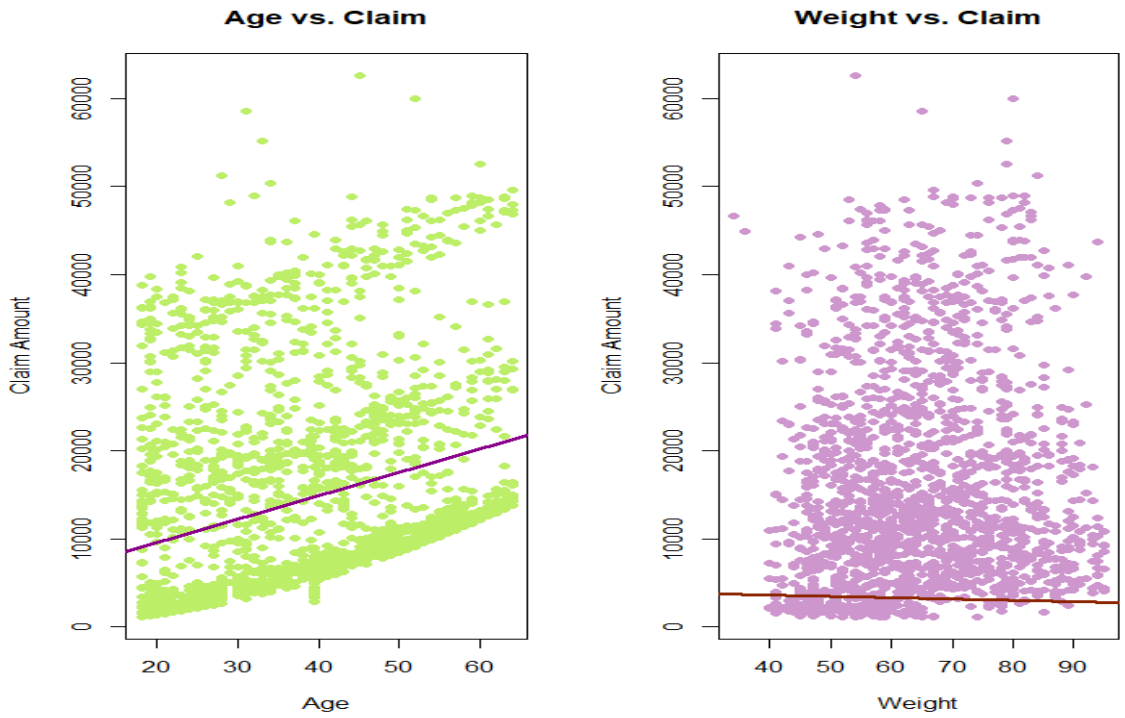
   In examining the influence of age compared to other factors like weight and exercise in predicting health insurance claim amounts, we employed linear regression. The primary focus was on understanding the relationships between the response variable (claim amounts) and the predictors: age, weight, and regular exercise status. Through the linear regression model, we extracted insights from the coefficients, shedding light on the magnitude and significance of each predictor's impact on health insurance claims. To enhance comprehension, separate plots were crafted to visually showcase the relationships between age and claims, as well as between weight and claims, providing a comprehensive comparison of their respective influences on the prediction of health insurance claim amounts.



Scatterplot of Age v/s Health Insurance Claim Amounts with Regular Exercise

```
> #1.   How does age compare to other factors like weight or exercise in predicting healt
h insurance claim amounts?
> # Fit a linear regression model
> Linear_regression_model <- lm(claim ~ age + weight + regular_exercise, data = final_dat
aset)
> # Summarize the model
> summary(Linear_regression_model)

Call:
lm(formula = claim ~ age + weight + regular_exercise, data = final_dataset)

Residuals:
   Min     1Q Median     3Q    Max
-10891  -6898  -5759   4921  47168

Coefficients:
                  Estimate Std. Error t value            Pr(>|t|)
(Intercept)       4258.266    497.544   8.559 <0.0000000000000002 ***
age                266.485      7.262  36.696 <0.0000000000000002 ***
weight             -15.284      7.455  -2.050               0.0404 *
regular_exercise -2069.720    233.443  -8.866 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11460 on 13963 degrees of freedom
Multiple R-squared:  0.0969,    Adjusted R-squared:  0.0967
F-statistic: 499.4 on 3 and 13963 DF,  p-value: < 0.00000000000000022
```

In the linear regression model predicting the insurance claim amount based on age, weight, and regular exercise, several key insights emerge. Age demonstrates a positive relationship with claim amount, with each one-unit increase in age associated with an expected increase in claim amount by 266.485 units, all else being constant. Conversely, weight exhibits a negative association, where a one-unit increase in weight corresponds to a decrease in claim amount by 15.284 units, holding other variables constant. Notably, engaging in regular exercise is linked to a substantial decrease in claim amount, with those who exercise regularly expected to have claim amounts lower by 2069.720 units compared to non-exercisers, controlling for other factors. The R-squared value of 0.0969 indicates that approximately 9.69% of the variance in the claim amount can be explained by age, weight, and regular exercise. Furthermore, the adjusted R-squared value of 0.0967, accounting for the number of predictors, suggests that the model's explanatory power remains largely unchanged. The F-statistic of 499.4, coupled with a very low p-value, underscores the model's overall significance, indicating that at least one of the predictor variables significantly contributes to predicting claim amount.

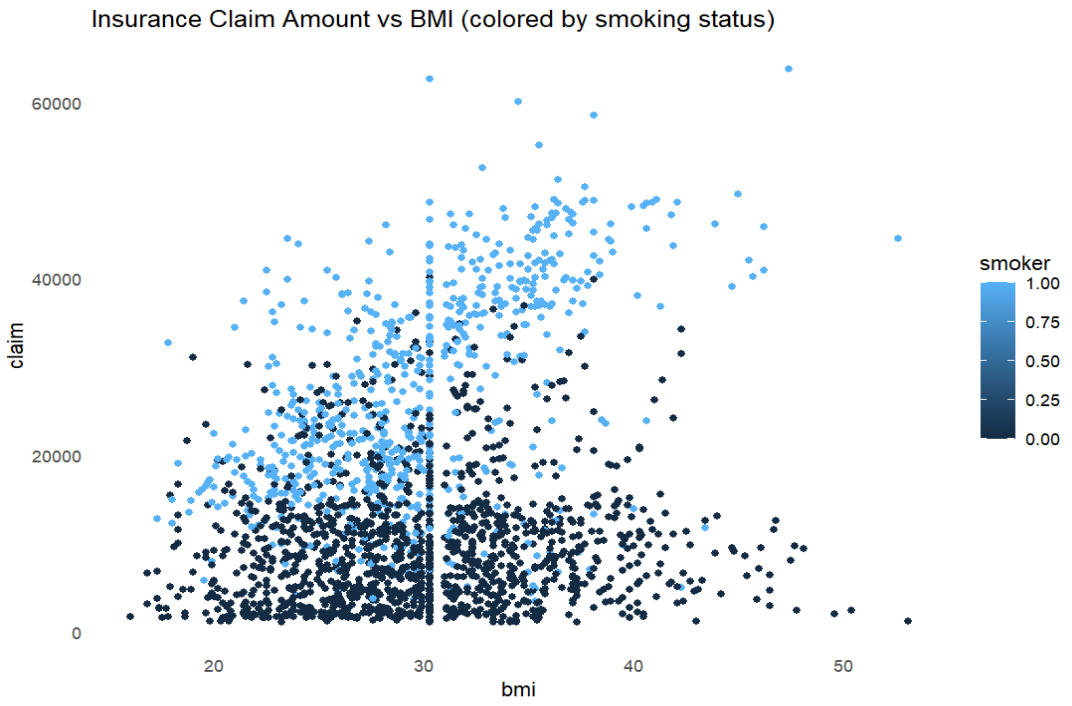


Linear Regression Graph for Age v/s Claim and Weight v/s Claim

2. Can we predict the likelihood of an individual having diabetes based on age and BMI?

In tackling the task of predicting diabetes likelihood based on age and BMI, logistic regression was employed with 'diabetes' as the binary response variable. Age and BMI served as the predictors in the model. The logistic regression model's summary provided insights into the impact and significance of age and BMI on the probability of diabetes.

To assess the model's predictive capabilities, we computed individual probabilities of having diabetes using age and BMI. The Receiver Operating Characteristic (ROC) curve visually depicted the model's discrimination ability, while the Area Under the Curve (AUC) quantified its overall accuracy. The logistic regression model, utilizing age and BMI as predictors, offers a means to estimate the likelihood of diabetes, contributing to a comprehensive understanding of the factors influencing this health outcome

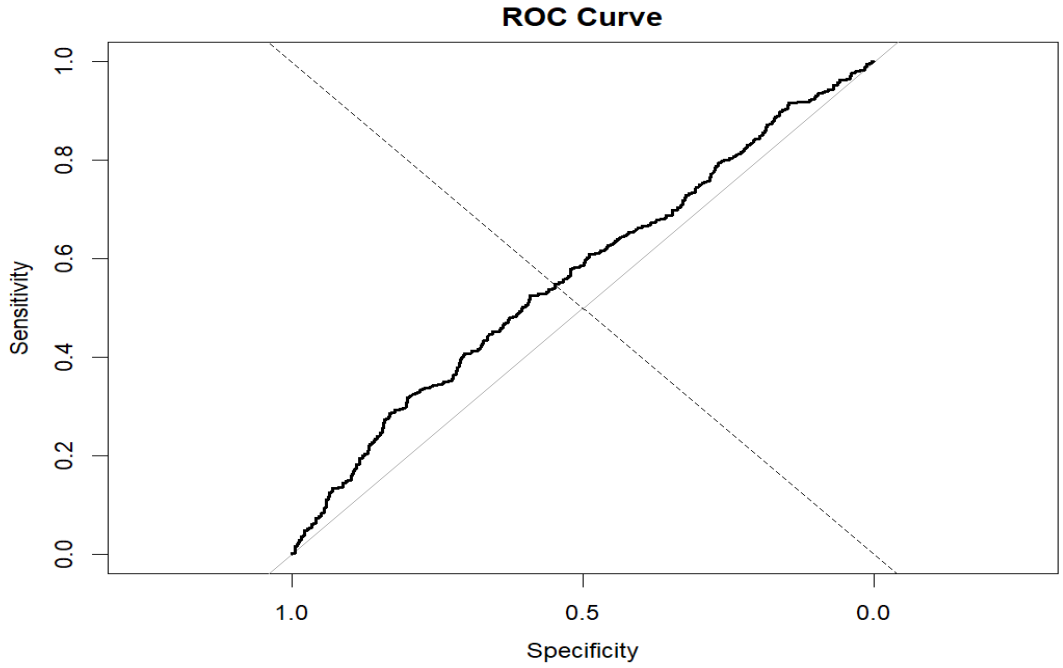


Scatterplot of BMI v/s Claim by Smoking Status

```
> # Fit a logistic regression model
> logistic_model <- glm(diabetes ~ age + bmi, data = final_dataset, family = "binomial")
> # Summarize the model
> summary(logistic_model)

Call:
glm(formula = diabetes ~ age + bmi, family = "binomial", data = final_dataset)

Coefficients:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept) -0.067219   0.115741  -0.581                0.561
age          0.008379   0.001496   5.601        0.0000000214 ***
bmi          0.032981   0.003730   8.842 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14877  on 13966  degrees of freedom
Residual deviance: 14742  on 13964  degrees of freedom
AIC: 14748

Number of Fisher Scoring iterations: 4
```

In the logistic regression model predicting the likelihood of having diabetes based on age and BMI, the coefficients offer valuable insights. A one-unit increase in age corresponds to a 0.008379 increase in the log-odds of having diabetes, while a similar increase in BMI results in a higher increment of 0.032981 in the log-odds, holding other variables constant. The intercept represents the log-odds of having diabetes when both age and BMI are zero, but its interpretation may not hold significance in this model's context. The positive

coefficients for age and BMI signify that as these variables increase, the log-odds of having diabetes also increase. Assessing goodness of fit, the null deviance of 14877 and residual deviance of 14742 indicate that the model improves upon the null model, suggesting a better fit as evidenced by the smaller residual deviance.



ROC Curve for logistic regression

```
> # Area under the curve
> Auc_curve <- auc(roc_obj)
> # Display the Auc curve
> Auc_curve
Area under the curve: 0.5659
```

The Area Under the Curve (AUC) value of 0.5659 indicates the discriminatory power of a binary classification model, commonly assessed through a Receiver Operating Characteristic (ROC) curve. An AUC around 0.5 suggests modest performance, only slightly better than random chance. A higher AUC is generally desired, indicating improved ability to distinguish between positive and negative classes.

3. How do the variables relate to smoking impact health insurance claims?

We explored the impact of various predictors, including smoking status, on health insurance claims using a systematic approach. The dataset underwent a 70-30 split into training and testing sets, with health insurance claim amount as the response variable and

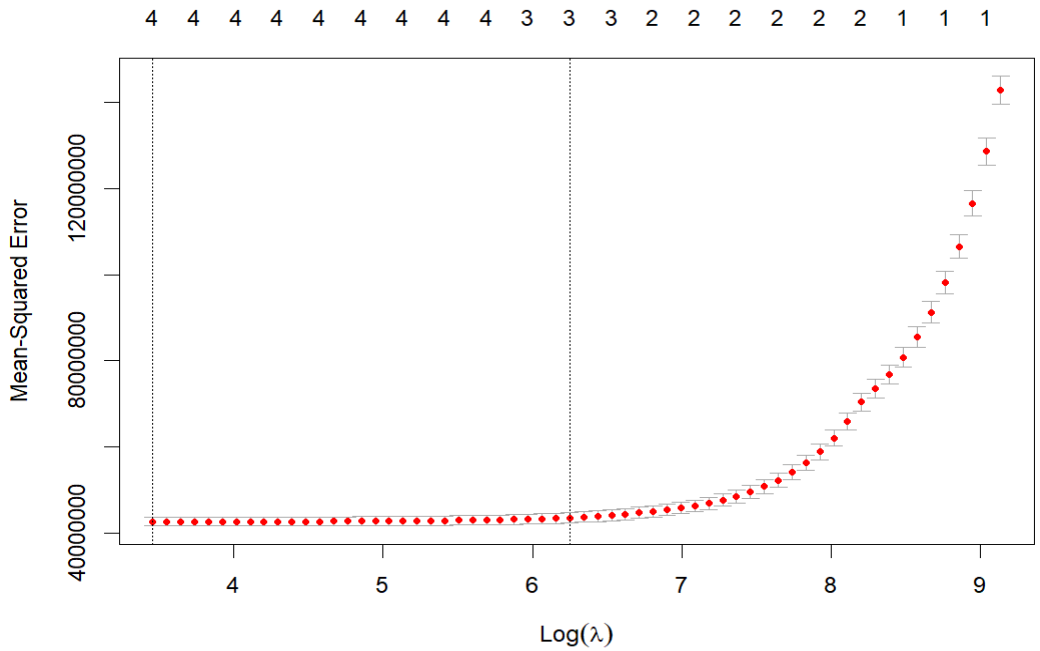
predictors such as smoking status, age, weight, and diabetes. Employing Lasso regression, a regularization technique, we created a sparse and interpretable model by penalizing less influential predictors. The Lasso model was trained on the subset of the data, and the model matrix, lacking an intercept term, was prepared. Optimal lambda, a regularization parameter, was chosen through cross-validation.

A coefficients plot visualized the impact of each predictor on health insurance claims, highlighting significant variables and those subject to regularization. To assess the model's accuracy on new data, we calculated the Root Mean Squared Error (RMSE). This comprehensive analysis allowed us to understand how predictors, especially smoking status, contribute to variations in health insurance claims.

```
> # Split the data into training and testing sets (e.g., 80-20 split)
> set.seed(123)  # for reproducibility
> train_index <- sample(seq_len(nrow(final_dataset)), 0.7 * nrow(final_dataset))
> train_data <- final_dataset[train_index, ]
> test_data <- final_dataset[-train_index, ]
> # Prepare the training data without intercept
> X_train <- model.matrix(claim ~ smoker + age + weight + diabetes - 1, data = train_dat
a)
> y_train <- train_data$claim
> # Fit Lasso regression model on training data
> lasso_model <- cv.glmnet(X_train, y_train, alpha = 1)
> print(lasso_model)

Call:  cv.glmnet(x = X_train, y = y_train, alpha = 1)

Measure: Mean-Squared Error

    Lambda Index  Measure      SE Nonzero
min   31.8    62 42660529 1034210       4
1se  517.6    32 43556285 1184366       3
> # Plot the coefficients
> plot(lasso_model)
> # Prepare the testing data without intercept
> X_test <- model.matrix(claim ~ smoker + age + weight + regular_exercise - 1, data = tes
t_data)
> y_test <- test_data$claim
> # Predict on the testing data
> predictions <- predict(lasso_model, newx = X_test, s = lasso_model$lambda.min)
> # Calculate RMSE
> rmse_lasso <- sqrt(mean((predictions - y_test)^2))
> print(paste("RMSE:", rmse_lasso))
[1] "RMSE: 6936.76767522832"
```

The output is from cross-validated LASSO regression using the cv.glmnet function, measuring the Mean Squared Error (MSE) for different values of the regularization parameter (lambda).

For the minimum MSE, the chosen lambda is 31.8, resulting in a mean squared error of 42,660,529 with 4 non-zero coefficients.
The 1-standard-error rule selects a lambda of 517.6, yielding a mean squared error of 43,556,285 with 3 non-zero coefficients.
The "RMSE: 6936.77" indicates the Root Mean Squared Error associated with the selected lambda.
.



Lasso Regression Plot

The image shows a plot typically used in regularization techniques like lasso regression. This plot displays the relationship between log(lambda) values on the x-axis and the mean squared error (MSE) on the y-axis. As lambda increases, so does the MSE. The red dots represent the MSE for each lambda value during cross-validation, and the vertical bars indicate the variability of the MSE. The lowest point on the curve represents the lambda value that minimizes the cross-validated MSE, often chosen for the final model to balance between model complexity and predictive power.

4. **How does the inclusion of regular exercise and blood pressure impact health insurance claims?**

We addressed the impact of regular exercise and blood pressure on health insurance claims by employing Ridge regression. The dataset was divided into training (70%) and testing (30%) sets. The response variable was the health insurance claim amount, and the predictors were 'regular_exercise' and 'bloodpressure.' The training set was utilized to construct a model matrix without an intercept, featuring the selected predictors. Ridge regression was applied using the cv.glmnet function with alpha set to 0 for Ridge regularization. Cross-validation determined the optimal lambda, and we examined the Ridge model by plotting its coefficients. Our goal was to assess how regular exercise and blood pressure influence health insurance claims. Predictions were made on the testing set using the Ridge model and the optimal lambda. The Root Mean Squared Error (RMSE) served as a metric to quantify the model's accuracy in predicting health insurance claims on new data.

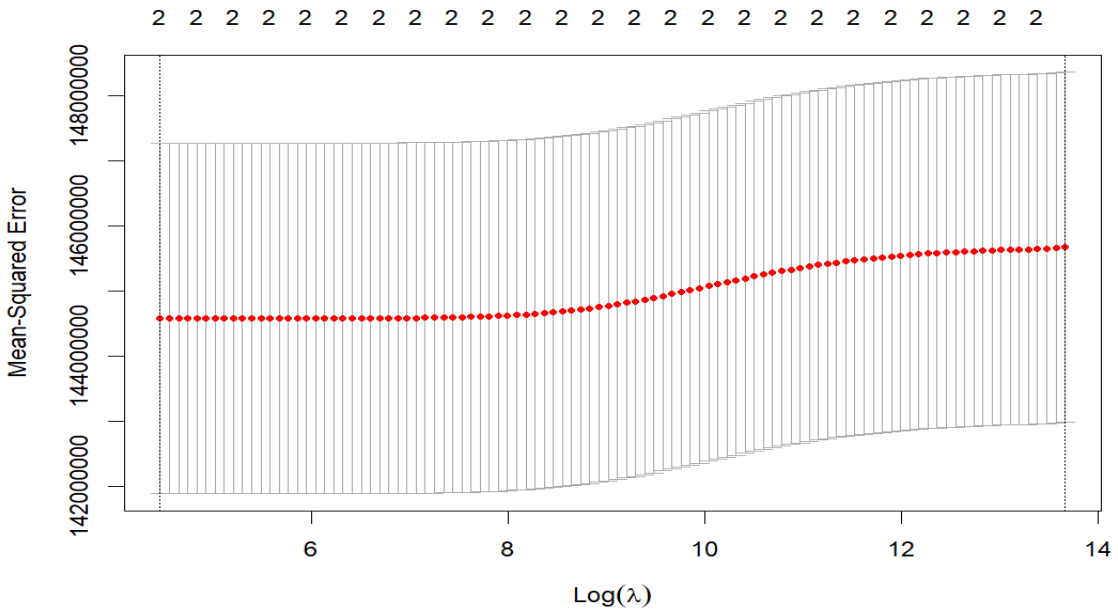
```
> # Split the data into training and testing sets
> set.seed(123)
> index <- createDataPartition(final_dataset$claim, p = 0.7, list = FALSE)
> train_data <- final_dataset[index, ]
> test_data <- final_dataset[-index, ]
> # Prepare the training data without intercept
> X_train_ridge <- model.matrix(claim ~ regular_exercise + bloodpressure - 1, data = trai
n_data)
> y_train_ridge <- train_data$claim
> # Fit Ridge regression model
> ridge_model <- cv.glmnet(X_train_ridge, y_train_ridge, alpha = 0)  # alpha = 0 for Ridg
e
> print(ridge_model)

Call:  cv.glmnet(x = X_train_ridge, y = y_train_ridge, alpha = 0)

Measure: Mean-Squared Error

    Lambda Index   Measure      SE Nonzero
min     86   100 144581667 2689794       2
1se 860464     1 145679428 2685077       2
> # Plot the coefficients
> plot(ridge_model)
> # Display the optimal lambda value chosen by cross-validation
> print(ridge_model$lambda.min)
[1] 86.0464
> # Prepare the testing data without intercept
> X_test_ridge <- model.matrix(claim ~ regular_exercise + bloodpressure - 1, data = test_
data)
> y_test_ridge <- test_data$claim
> # Make predictions
> predictions <- predict(ridge_model, s = ridge_model$lambda.min, newx = X_test_ridge)
> # Calculate RMSE
> rmse_ridge <- sqrt(mean((predictions - y_test_ridge)^2))
> print(paste("RMSE:", rmse_ridge))
[1] "RMSE: 11989.5261133348"
```

The fitting of a Ridge regression model using cross-validation to determine the optimal regularization parameter (lambda). This parameter helps control overfitting by penalizing large coefficients in the regression model.

After fitting the model, the cross-validation process selects the lambda value (**86.0464**) that minimizes the mean-squared error. The plotted coefficients illustrate the effect of regularization on the model parameters.

Subsequently, predictions are generated on the testing dataset, and the Root Mean Squared Error (RMSE) is calculated to evaluate the model's performance. In this case, the RMSE value of **11989.526** indicates the average difference between the actual and predicted claim amounts in the testing dataset. Lower RMSE values generally suggest better model performance.

While the RMSE suggests reasonable predictive capability, the model's effectiveness should be further assessed within the context of the specific problem domain. Comparing its performance with alternative models or benchmarks would provide a clearer understanding of its relative strengths and weaknesses.



Ridge Regression Plot

The image displays a ridge trace plot from a ridge regression model, as indicated by the pattern of the mean squared error (MSE) across different log(lambda) values. This type of plot helps in selecting the regularization parameter (lambda) for the ridge regression. The red dots denote the MSE for each lambda value tested. The plot shows that as log(lambda) increases, the MSE remains relatively stable after a certain point, suggesting that further increases in lambda do not significantly increase the error. This stability implies that an optimal lambda value might be found within the plateau region of the plot.

Conclusion and Project Goals

In conclusion, the analysis of the Health Insurance Dataset offers a comprehensive understanding of the intricate factors influencing health insurance claims, vital for enhancing risk assessment accuracy and ensuring equitable coverage for policyholders. Through a systematic exploration of various dataset attributes, ranging from demographic details like age and gender to lifestyle indicators such as smoking and exercise habits, the study sheds light on critical insights essential for insurers and policymakers alike.

The identification and handling of missing values, particularly in key variables like age and BMI, underlines the importance of data integrity and robust preprocessing techniques. By employing methods like imputation with mean values, the dataset's completeness and representativeness are preserved, laying a solid foundation for subsequent analyses.

Correlation analysis uncovers significant associations among variables, revealing nuanced relationships that inform risk assessment strategies and premium calculations. Noteworthy findings include the positive correlation of age with weight and BMI, as well as the strong correlation between smoking and higher insurance claims, emphasizing the multifaceted nature of risk factors.

Furthermore, outlier detection and removal using the Interquartile Range (IQR) method ensures that extreme values do not unduly influence subsequent analyses and visualizations, thereby enhancing the reliability of insights derived from the dataset.

The transformation of categorical data into numerical representations facilitates quantitative analysis, enabling deeper insights into the interplay between various factors and their impact on insurance claims. Visualizations such as histograms and bar plots provide intuitive insights into claim distribution patterns, highlighting notable trends such as higher claim amounts for non-smokers and individuals with no regular exercise.

Regression analysis, including linear and logistic regression models, enables the prediction of health insurance claim amounts and the likelihood of diabetes based on key factors like age, BMI, and lifestyle choices. Regularization techniques such as Ridge and Lasso regression mitigate overfitting and enhance model generalizability, contributing to more accurate risk profiling and policy pricing.

Reference

https://www.kaggle.com/datasets/sureshgupta/health-insurance-data-set