Predicting Heart Attack Risk Through Data Analysis and Machine Learning Models

Nana Firdausi Hassan

Table of Contents

THE DATA

Description of Data

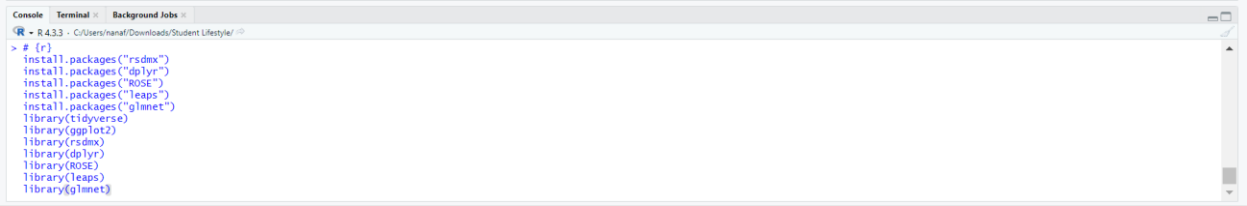
The Heart Attack Prediction Dataset was obtained from Kaggle, it contains detailed information about patients and their risk of heart attack. The dataset includes 8,763 observations and 26 variables, offering a wealth of data on patient demographics, medical history, and diagnostic measurements.

This dataset provides valuable resources for data scientists and healthcare researchers who aim to explore the factors contributing to heart attack risk and develop machine learning models for early detection and prevention.

Link to the dataset: [Heart Attack Prediction Dataset on Kaggle]

Loading Packages

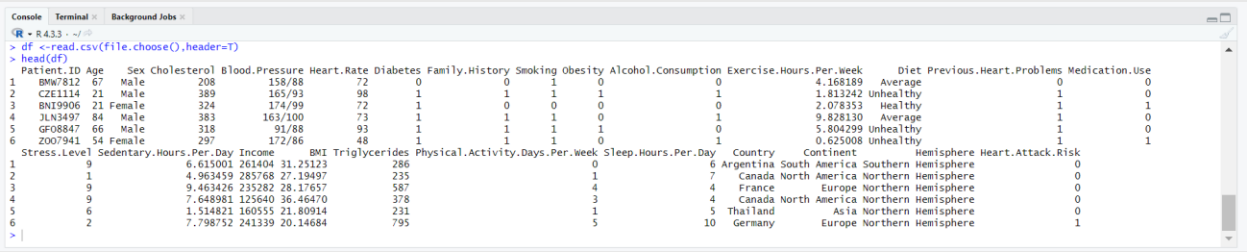
For this analysis, I installed and loaded several R packages to support data processing, visualization, and modeling. dplyr and tidyverse were used for data manipulation and cleaning, while ggplot2 supported data visualization. The ROSE package helped address class imbalance in the dataset. For variable selection and model building, I used leaps for subset selection and glmnet for regularized regression techniques like Lasso and Ridge. Additionally, rsdmx was included to allow access to structured statistical data formats when needed. More packages will be added as the analysis progresses and additional functionality is required.



```
> # {r}
  install.packages("rsdmx")
  install.packages("dplyr")
  install.packages("ROSE")
  install.packages("leaps")
  install.packages("glmnet")
  library(tidyverse)
  library(ggplot2)
  library(rsdmx)
  library(dplyr)
  library(ROSE)
  library(leaps)
  library(glmnet)
```

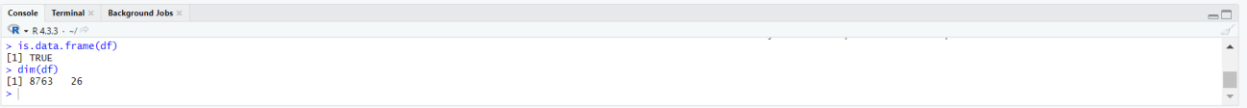
Loading Dataset

I imported the dataset using read.csv() and used head() to view the first few rows. This step helped confirm the data loaded correctly and provided an overview of the variables for further analysis.



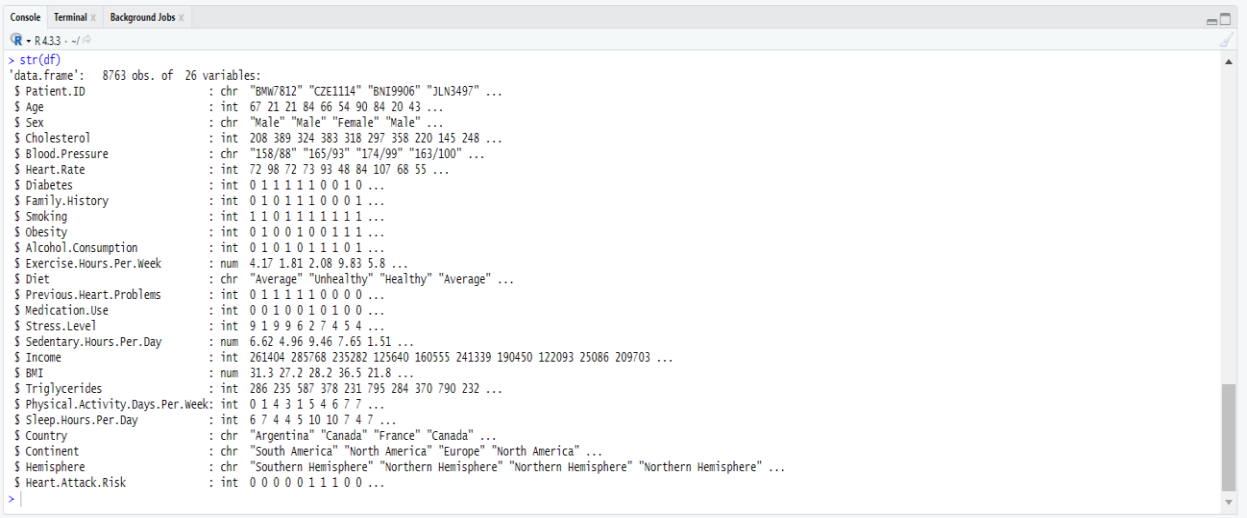
```
> df <-read.csv(file.choose(),header=T)
> head(df)
  Patient.ID Age    Sex Cholesterol Blood.Pressure Heart.Rate Diabetes Family.History Smoking Obesity Alcohol.Consumption Exercise.Hours.Per.Week      Diet Previous.Heart.Problems Medication.Use
1    BMW7812  67   Male         208        158/88         72        0              0       1       0                   0              4.168189   Average                        0              0
2    CZE1114  21   Male         389        165/93         98        1              1       1       1                   1              1.813242 Unhealthy                        1              0
3    BNI9906  21 Female         324        174/99         72        1              0       0       0                   0              2.078353   Healthy                        1              1
4    JLN3497  84   Male         383        163/100        73        1              1       1       0                   1              9.828130   Average                        1              0
5    GFO8847  66   Male         318         91/88         93        1              1       1       1                   0              5.804299 Unhealthy                        1              0
6    ZOO7941  54 Female         297        172/86         48        1              1       1       0                   1              0.625008 Unhealthy                        1              1
  Stress.Level Sedentary.Hours.Per.Day Income      BMI Triglycerides Physical.Activity.Days.Per.Week Sleep.Hours.Per.Day   Country     Continent          Hemisphere Heart.Attack.Risk
1            9              6.615001 261404 31.25123          286                               0                   6 Argentina South America Southern Hemisphere                0
2            1              4.963459 285768 27.19497          235                               1                   7    Canada North America Northern Hemisphere                0
3            9              9.463426 235282 28.17657          587                               4                   4    France        Europe Northern Hemisphere                0
4            9              7.648981 125640 36.46470          378                               3                   4    Canada North America Northern Hemisphere                0
5            6              1.514821 160555 21.80914          231                               1                   5  Thailand        Asia Northern Hemisphere                0
6            2              7.798752 241339 20.14684          795                               5                  10   Germany        Europe Northern Hemisphere                1
>
```

I used is.data.frame(df) to confirm the dataset is a data frame and is correctly loaded into R, which returned TRUE. Then, I checked the dimensions with dim(df), revealing that the dataset has 8763 observations and 26 variables.

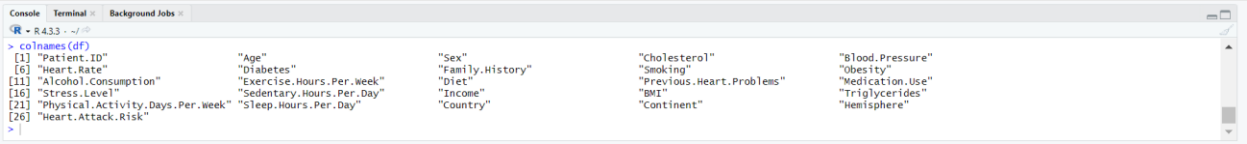


Original Data Structure

To obtain information about the variables in the dataset, I used the str(df) function, which displays the structure of the dataset, including its mode and data type. By using the str() function, I was able to determine that the dataset had variables with different modes, including numeric (num), integer (int), and character (chr) modes. The dataset includes both categorical variables and numerical variables.



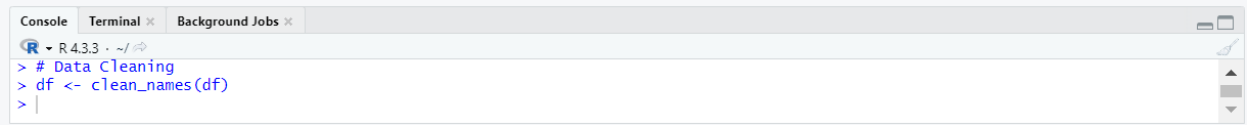
I used the colnames(df) function to retrieve the column names of the dataset.



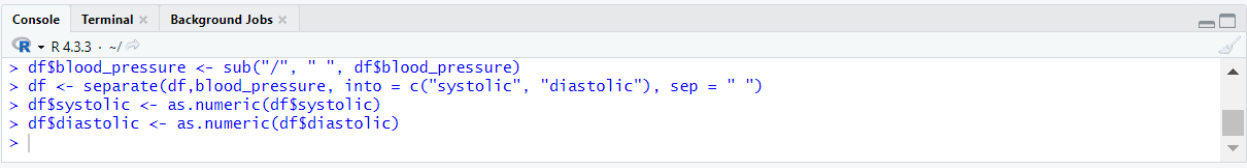
Data Cleaning

As the first step in my data cleaning process, I used the clean_names() function from the janitor package to standardize the column names in the dataset. This function reformats all column names to lowercase, replaces spaces and special characters with underscores, and ensures that names are unique and syntactically valid. For example, a

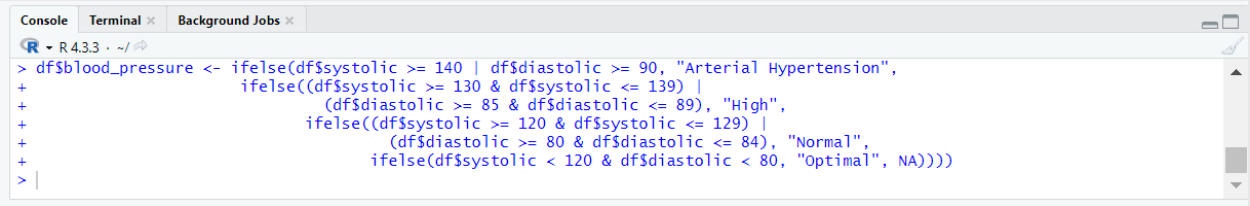
column like "Patient.ID" is converted to "patient_id". This step improves consistency, prevents potential coding errors, and makes data manipulation easier in the following stages of analysis.



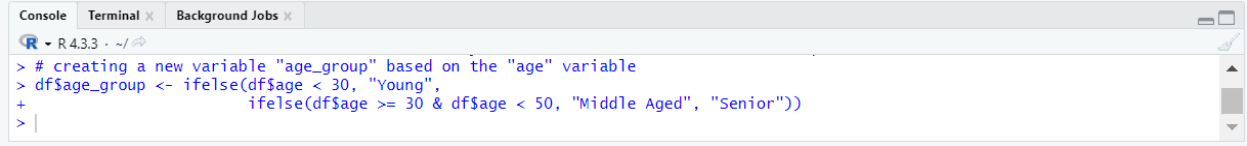
I cleaned the blood_pressure column by first replacing the "/" with a space using the sub() function. Then, I split the column into two separate columns, systolic and diastolic, using the separate() function. Finally, I converted both columns to numeric format with as.numeric() to allow for proper analysis. This step made it possible to work with systolic and diastolic values independently in numerical operations.



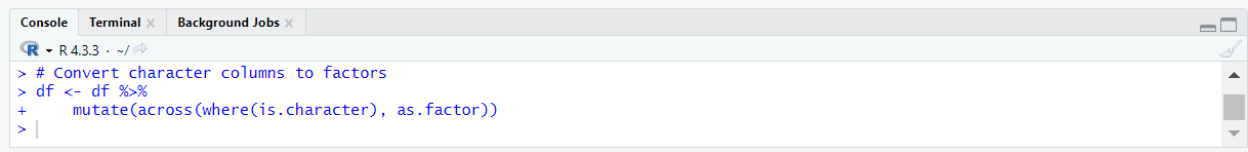
I created a new classification column for ‘blood pressure’ using nested ifelse() statements. Readings with systolic ≥ 140 or diastolic ≥ 90 were classified as "Arterial Hypertension." Values with systolic 130–139 or diastolic 85–89 were labeled "High," those with systolic 120–129 or diastolic 80–84 as "Normal," and readings with systolic < 120 and diastolic < 80 as "Optimal." This allowed for standardized categorization of blood pressure levels for analysis.



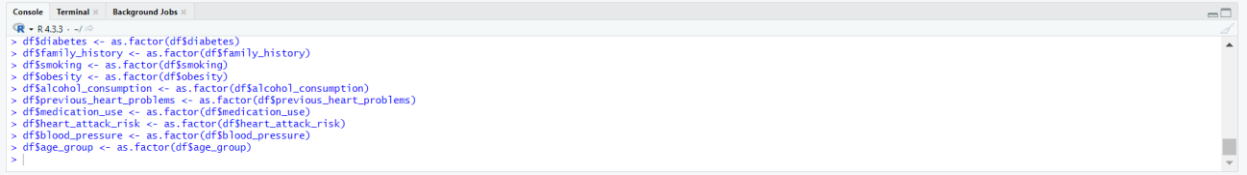
I created a new variable, age_group, to categorize individuals based on age: "Young" for ages under 30, "Middle Aged" for ages 30–49, and "Senior" for ages 50 and above.



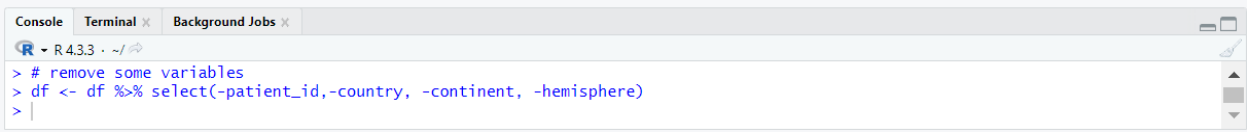
I used ‘mutate’ with across(where(is.character), as.factor) in R to convert all character columns in my dataset into factors. This transformation ensures that the dataset is optimized for statistical analysis and modeling.



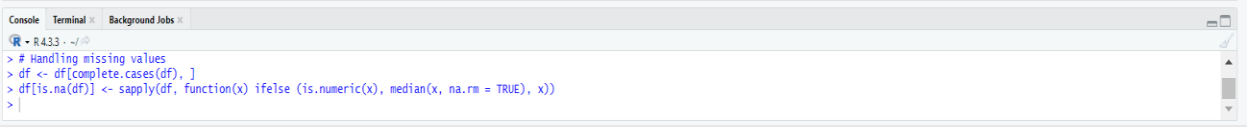
I converted relevant columns to factor type using as.factor() to ensure they are properly recognized as categorical variables during analysis and modeling. This step improves the accuracy of statistical summaries and machine learning algorithms that handle categorical data.



I removed non-essential columns such as identifiers and geographic details using the select() function with the minus operator. This helped streamline the dataset by keeping only relevant variables for analysis.



I addressed missing values in the dataset by first removing rows with incomplete cases using complete.cases(). Then, for any remaining NA values, I replaced them with the median for numeric variables, ensuring no data distortion, while non-numeric values were left unchanged.



Cleaned Data Frame

I checked for any remaining missing values in the dataset using colSums(sapply(df, is.na)). This confirmed that all NA values had been handled, ensuring the dataset is clean and ready for analysis.

```
> # Cleaned Data set
> colSums(sapply(df, is.na))
                    age                     sex             cholesterol                systolic               diastolic              heart_rate
                      0                       0                       0                       0                       0                       0
               diabetes          family_history                 smoking                 obesity     alcohol_consumption  exercise_hours_per_week
                      0                       0                       0                       0                       0                       0
                   diet previous_heart_problems          medication_use          stress_level  sedentary_hours_per_day                  income
                      0                       0                       0                       0                       0                       0
                    bmi          triglycerides physical_activity_days_per_week  sleep_hours_per_day     heart_attack_risk        blood_pressure
                      0                       0                       0                       0                       0                       0
              age_group          family.history
                      0                       0
```

After cleaning the data, I used str(df) to check the structure of the dataset and ensure proper data types. I also reviewed the first few rows with head(df) to confirm that the data was correctly processed and formatted, making it ready for further analysis.

```
> str(df)
'data.frame':   8763 obs. of  25 variables:
 $ age                        : int  67 21 21 84 66 54 90 84 20 43 ...
 $ sex                        : Factor w/ 2 levels "Female","Male": 2 2 1 2 2 1 2 2 2 1 ...
 $ cholesterol                : int  208 389 324 383 318 297 358 220 145 248 ...
 $ systolic                   : num  158 165 174 163 91 172 102 131 144 160 ...
 $ diastolic                  : num  88 93 99 100 88 86 73 68 105 70 ...
 $ heart_rate                 : int  33 59 33 34 54 9 45 68 29 16 ...
 $ diabetes                   : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 1 1 2 1 ...
 $ family_history             : Factor w/ 2 levels "1","2": 1 2 1 2 2 2 1 1 1 2 ...
 $ smoking                    : Factor w/ 2 levels "0","1": 2 2 1 2 2 2 2 2 2 2 ...
 $ obesity                    : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 1 2 2 2 ...
 $ alcohol_consumption        : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 2 2 1 2 ...
 $ exercise_hours_per_week    : num  4.17 1.81 2.08 9.83 5.8 ...
 $ diet                       : Factor w/ 3 levels "Average","Healthy",..: 1 3 2 1 3 3 2 1 1 3 ...
 $ previous_heart_problems    : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 1 1 1 1 ...
 $ medication_use             : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 1 ...
 $ stress_level               : int  9 1 9 9 6 2 7 4 5 4 ...
 $ sedentary_hours_per_day    : num  6.62 4.96 9.46 7.65 1.51 ...
 $ income                     : int  261404 285768 235282 125640 160555 241339 190450 122093 25086 209703 ...
 $ bmi                        : num  31.3 27.2 28.2 36.5 21.8 ...
 $ triglycerides              : int  286 235 587 378 231 795 284 370 790 232 ...
 $ physical_activity_days_per_week: int  0 1 4 3 1 5 4 6 7 7 ...
 $ sleep_hours_per_day        : int  6 7 4 4 5 10 10 7 4 7 ...
 $ heart_attack_risk          : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 2 1 1 ...
 $ blood_pressure             : Factor w/ 4 levels "Arterial Hypertension",..: 1 1 1 1 2 1 4 2 1 1 ...
 $ age_group                  : Factor w/ 3 levels "Middle Aged",..: 2 3 3 2 2 2 2 2 3 1 ...
```

```
> head(df)
  age    sex cholesterol systolic diastolic heart_rate diabetes family_history smoking obesity alcohol_consumption exercise_hours_per_week      diet previous_heart_problems medication_use
1  67   Male         208      158        88         33        0              1       1       0                   0                 4.168189   Average                        0              0
2  21   Male         389      165        93         59        1              2       1       1                   1                 1.813242 Unhealthy                        1              0
3  21 Female         324      174        99         33        1              1       0       0                   0                 2.078353   Healthy                        1              1
4  84   Male         383      163       100         34        1              2       1       0                   1                 9.828130   Average                        1              0
5  66   Male         318       91        88         54        1              2       1       1                   0                 5.804299 Unhealthy                        1              0
6  54 Female         297      172        86          9        1              2       1       0                   1                 0.625008 Unhealthy                        1              1
  stress_level sedentary_hours_per_day income      bmi triglycerides physical_activity_days_per_week sleep_hours_per_day heart_attack_risk        blood_pressure age_group
1            9              6.615001 261404 31.25123          286                                 0                   6                 0 Arterial Hypertension    Senior
2            1              4.963459 285768 27.19497          235                                 1                   7                 0 Arterial Hypertension     Young
3            9              9.463426 235282 28.17657          587                                 4                   4                 0 Arterial Hypertension     Young
4            9              7.648981 125640 36.46470          378                                 3                   4                 0 Arterial Hypertension    Senior
5            6              1.514821 160555 21.80914          231                                 1                   5                 0                   High    Senior
6            2              7.798752 241339 20.14684          795                                 5                  10                 1 Arterial Hypertension    Senior
```

Description of Variables

This table outlines the variables in the "Heart Attack Risk Prediction Dataset. It provides descriptions and data types for each variable.

Variable	Description	Class
Patient ID	Unique identifier for each patient	Character
Age	Age of the patient	Integer
Sex	Gender of the patient (Male/Female)	Character
Cholesterol	Cholesterol levels of the patient	Integer
Blood Pressure	Blood pressure of the patient (systolic/diastolic)	Character
Heart Rate	Heart rate of the patient	Integer
Diabetes	Whether the patient has diabetes (Yes/No)	Integer

Family History	Family history of heart-related problems (1: Yes, 0: No)	Integer
Smoking	Smoking status of the patient (1: Smoker, 0: Non-smoker)	Integer
Obesity	Obesity status of the patient (1: Obese, 0: Not obese)	Integer
Alcohol Consumption	Level of alcohol consumption by the patient (None/Light/Moderate/Heavy)	Integer
Exercise Hours Per Week	Number of exercise hours per week	Numeric
Diet	Dietary habits of the patient (Healthy/Average/Unhealthy)	Character
Previous Heart Problems	Previous heart problems of the patient (1: Yes, 0: No)	Integer
Medication Use	Medication usage by the patient (1: Yes, 0: No)	Integer
Stress Level	Stress level reported by the patient (1-10)	Integer
Sedentary Hours Per Day	Hours of sedentary activity per day	Numeric
Income	Income level of the patient	Integer
BMI	Body Mass Index (BMI) of the patient	Numeric
Triglycerides	Triglyceride levels of the patient	Integer
Physical Activity Days Per Week	Days of physical activity per week	Integer
Sleep Hours Per Day	Hours of sleep per day	Integer
Country	Country of the patient	Character
Continent	Continent where the patient resides	Character
Hemisphere	Hemisphere where the patient resides	Character
Heart Attack Risk	Presence of heart attack risk (1: Yes, 0: No)	Integer

Expectations

This analysis aims to explore the heart attack risk dataset in detail, focusing on identifying the key variables that contribute to predicting heart attack risk. I will examine factors such as age, cholesterol levels, BMI, family history, blood pressure, alcohol consumption, diet, physical activity, etc. These variables will be evaluated using statistical methods, including logistic regression and classification analysis, to determine the most significant predictors of heart attack risk.

To enhance understanding, I will utilize data visualization techniques, such as scatter plots, box plots, and other charts, to identify patterns and trends within the data. These visualizations will aid in discovering the relationships between various health factors and the risk of heart attack.

The goal of this analysis is to generate actionable insights and recommendations for heart attack prevention. By accurately pinpointing the critical risk factors and creating reliable predictive models, the analysis will support the development of strategies to improve heart health and reduce the likelihood of heart attacks.

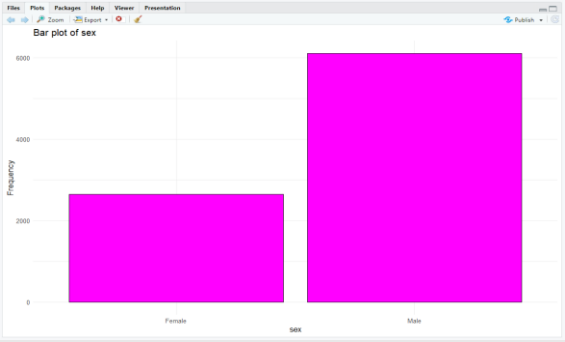
DATA EXPLORATION AND ANALYSIS

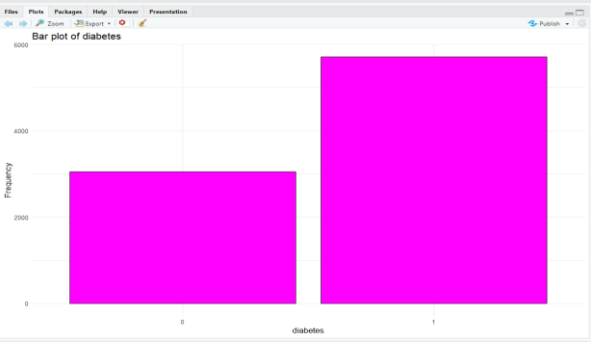
To begin my data exploration of the Heart Attack Risk Prediction Dataset, I used the summary() function to examine key variables such as age, cholesterol, heart rate, and diet. This provided basic statistical insights, including the 1st quartile, 3rd quartile minimum, maximum, median, and mean values for the numerical variables. For the categorical variable diet, I observed the frequency distribution across its categories: Average, Healthy, and Unhealthy. This initial analysis gave me a general understanding of the dataset's structure and helped identify patterns worth exploring further.

```
> #Data Analysis
> summary(select(df, age, cholesterol, heart_rate, diet))
      age         cholesterol      heart_rate          diet
 Min.   :18.00   Min.   :120.0   Min.   : 1.00   Average  :2912
 1st Qu.:35.00   1st Qu.:192.0   1st Qu.:18.00   Healthy  :2960
 Median :54.00   Median :259.0   Median :36.00   Unhealthy:2891
 Mean   :53.71   Mean   :259.9   Mean   :36.02
 3rd Qu.:72.00   3rd Qu.:330.0   3rd Qu.:54.00
 Max.   :90.00   Max.   :400.0   Max.   :71.00
>
```

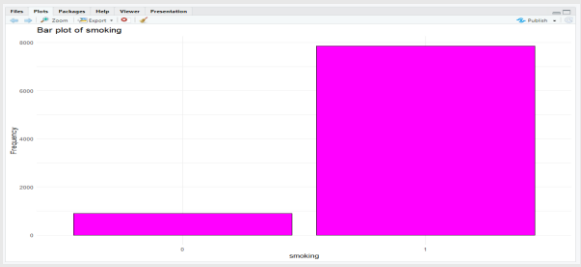
Plots for Categorical Variables

```
> # plot for categorical variables
> for (var_name in names(df)) {
+   if(is.factor(df[[var_name]])) {
+     #create a bar plot
+     p <- ggplot(df, aes_string(x = var_name)) +
+       geom_bar(fill = "magenta", color = "black") +
+       theme_minimal() +
+       labs(title = paste("Bar plot of", var_name), x =var_name, y ="Frequency")
+
+     print(p)
+   }
+
+ }
```

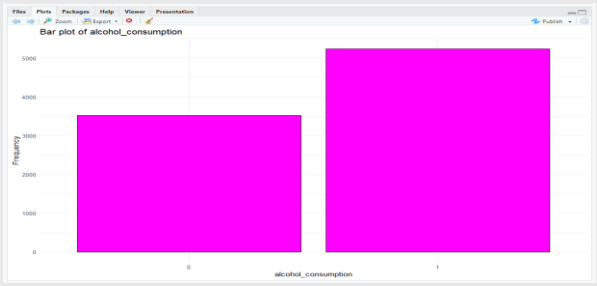


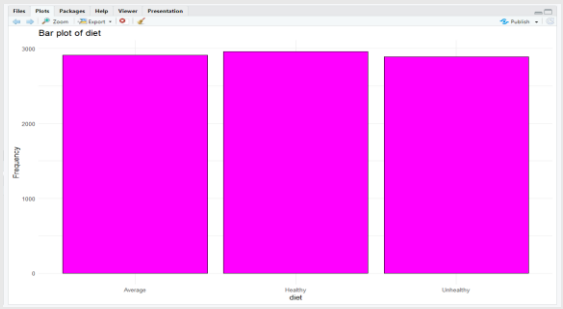


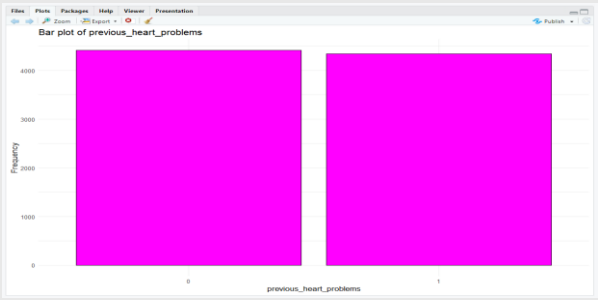

Bar plot of family_history
Frequency
family_history


Bar plot of smoking
Frequency
smoking

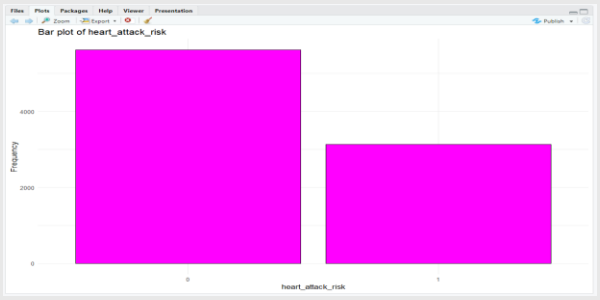

Bar plot of obesity
Frequency
obesity

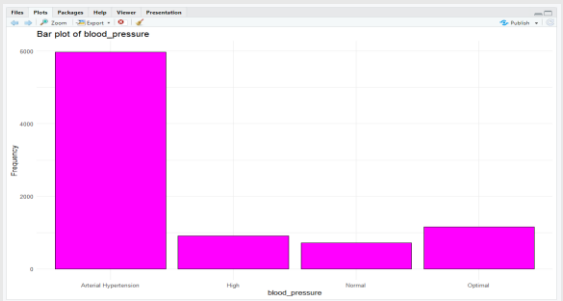

Bar plot of alcohol_consumption
Frequency
alcohol_consumption

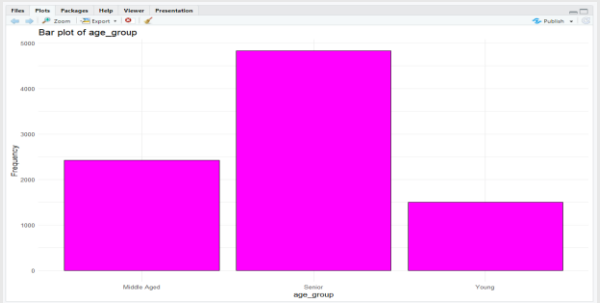

Bar plot of diet
Frequency
Average
Healthy
Unhealthy
diet


Bar plot of previous_heart_problems
Frequency
previous_heart_problems

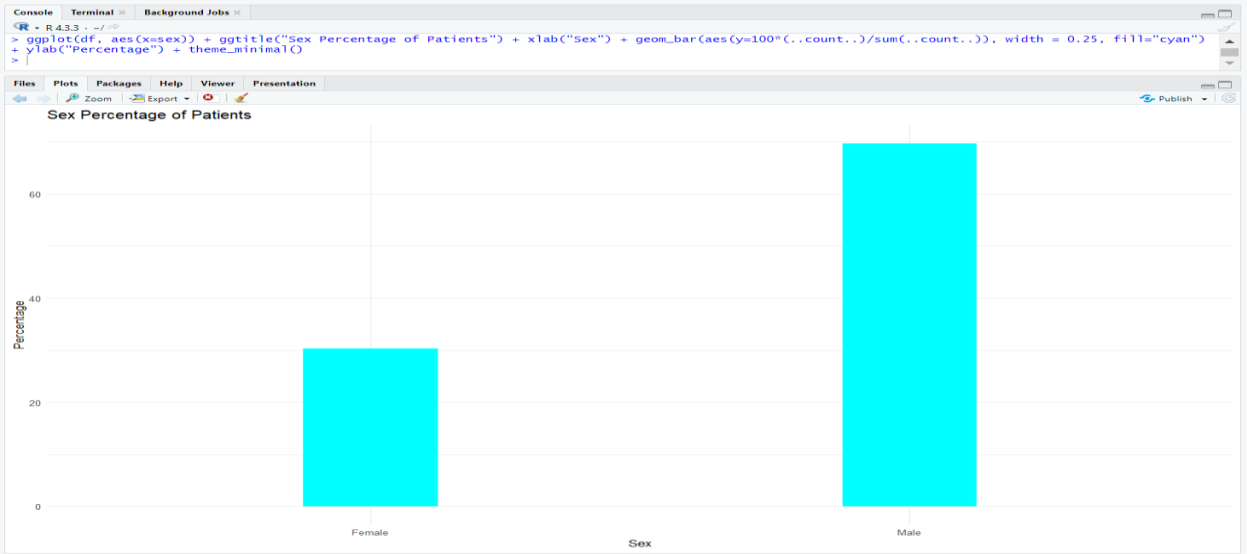

Bar plot of medication_use
Frequency
medication_use

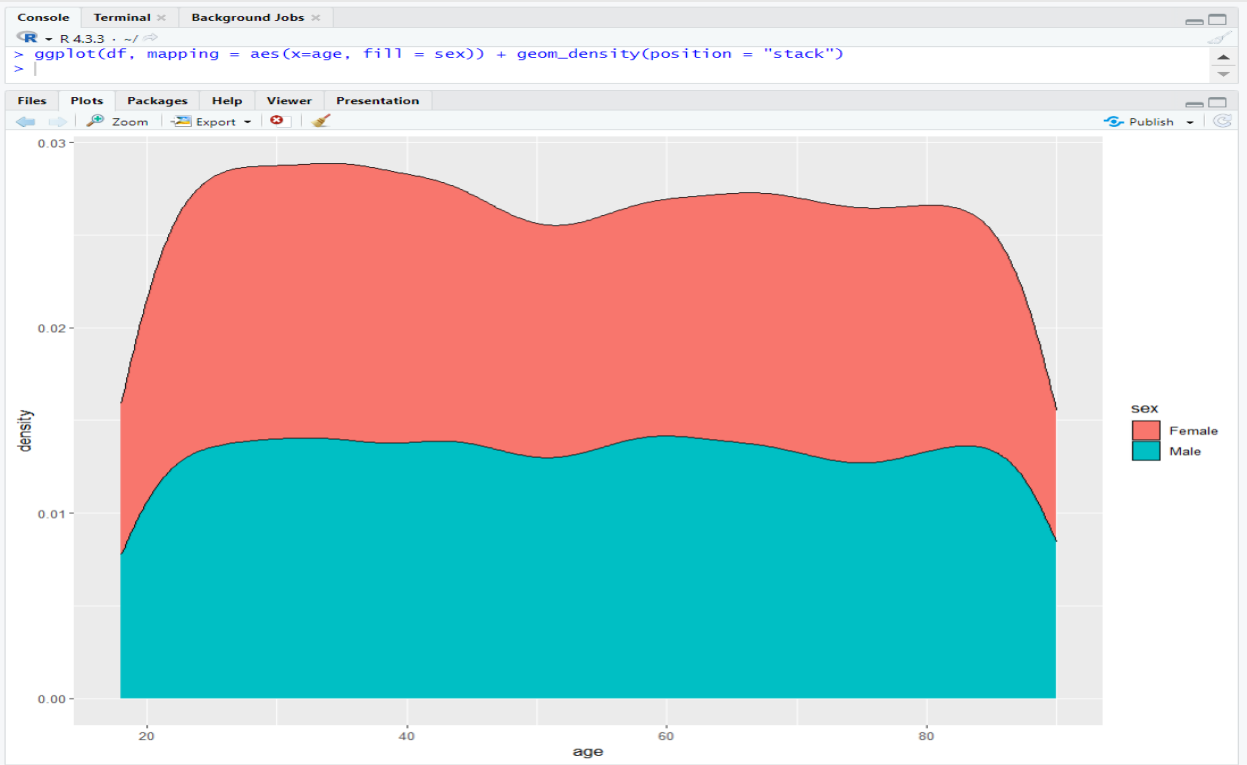

Bar plot of heart_attack_risk
Frequency
heart_attack_risk


Bar plot of blood_pressure
Frequency
Arterial Hypertension
High
Normal
Optimal
blood_pressure

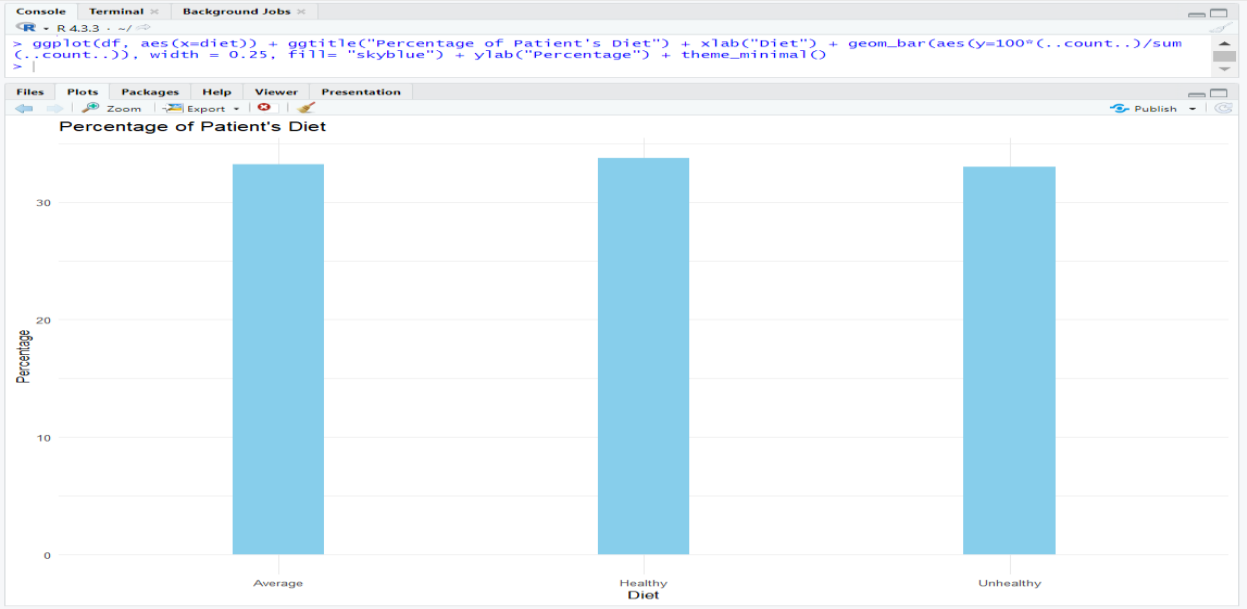

Bar plot of age_group
Frequency
Middle Aged
Senior
Young
age_group

To gain an insight into the sex distribution of patients, I generated a graph that depicts the percentage distribution in the ‘Heart Attack Risk Prediction Dataset’. The results reveal that male patients represent a larger percentage compared to female patients, highlighting a noticeable disparity in gender representation within the dataset. This observation underscores the predominance of male patients in the analyzed data



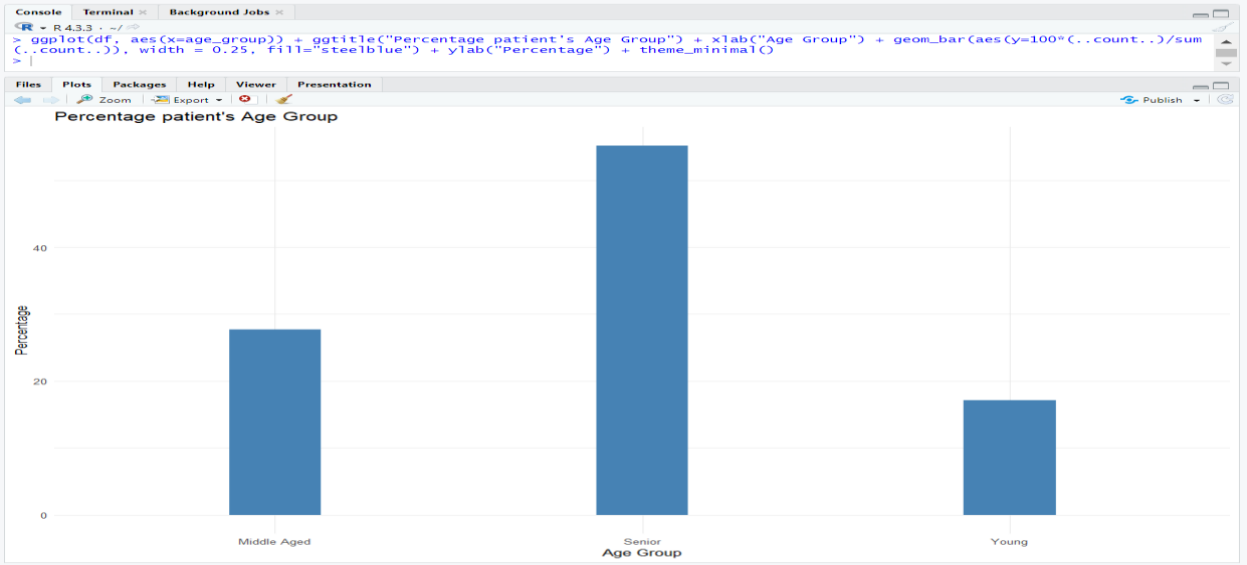


The bar chart below illustrates the percentage breakdown of a patient's dietary intake across three categories. Approximately 35% of the patient's diet consists of food classified as average. Similarly, healthy food choices also constitute about 35% of the patient's diet. The remaining portion, roughly 34%, is categorized as unhealthy. Therefore, the patient's diet is almost evenly split between average, healthy, and unhealthy food options.

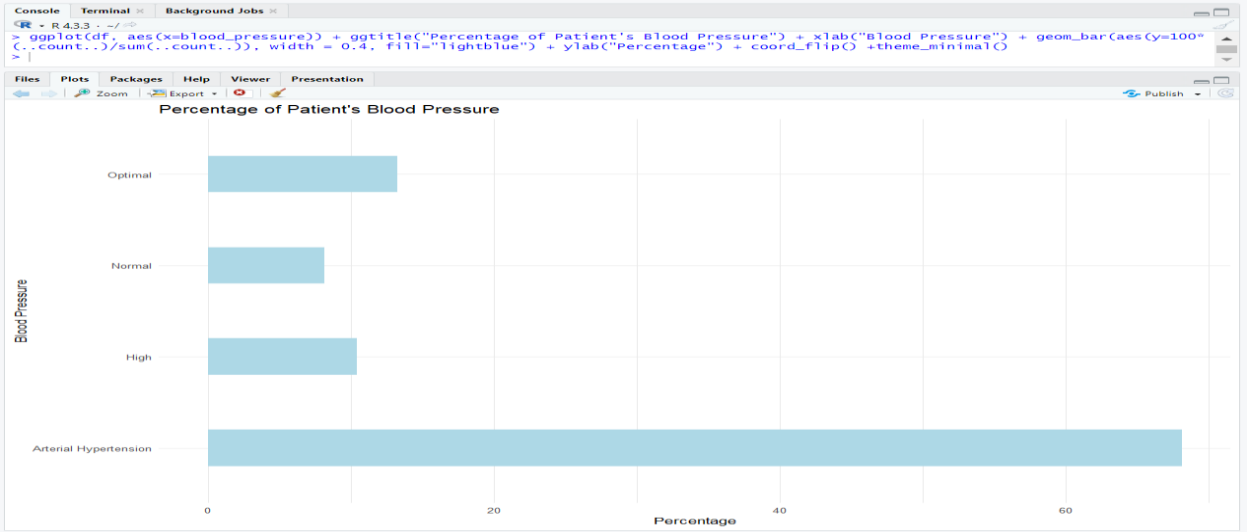


The bar chart shows the distribution of individuals across diet categories (Average, Healthy, Unhealthy), segmented by age group (Young, Middle Aged, Senior). In all diet categories, Seniors have the highest count, followed by Middle Aged individuals, while Young individuals have the lowest. This pattern is consistent across all diet groups.

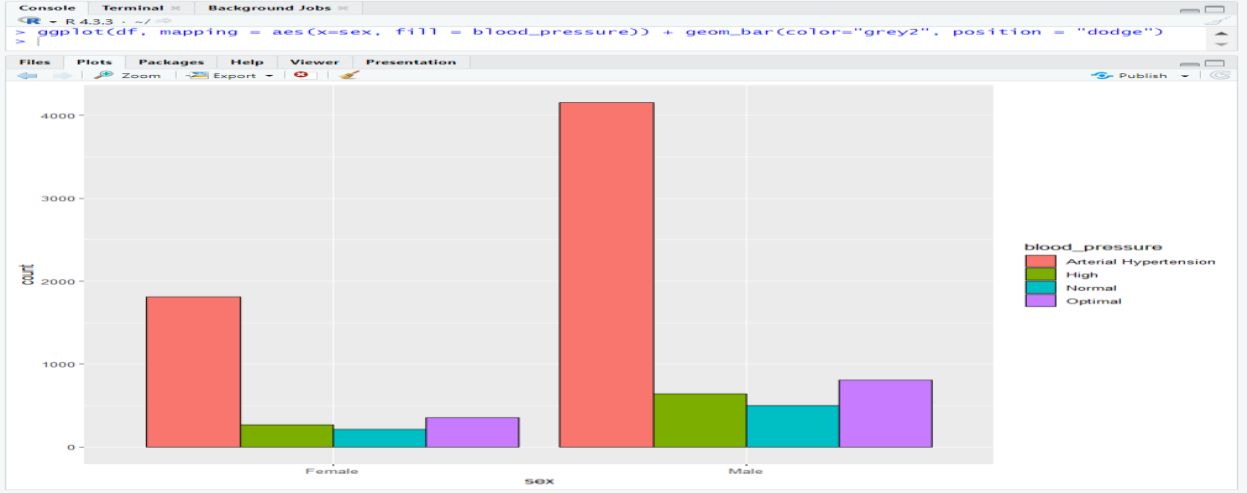
The bar chart displays the percentage distribution of patients across three age groups. The "Middle Aged" group represents approximately 28% of the patients. A significantly larger proportion, around 55%, falls into the "Senior" age group. The "Young" age group constitutes the smallest percentage, at approximately 17%. Thus, most of the patients in this dataset are seniors, followed by middle-aged individuals, with the smallest representation being young patients.



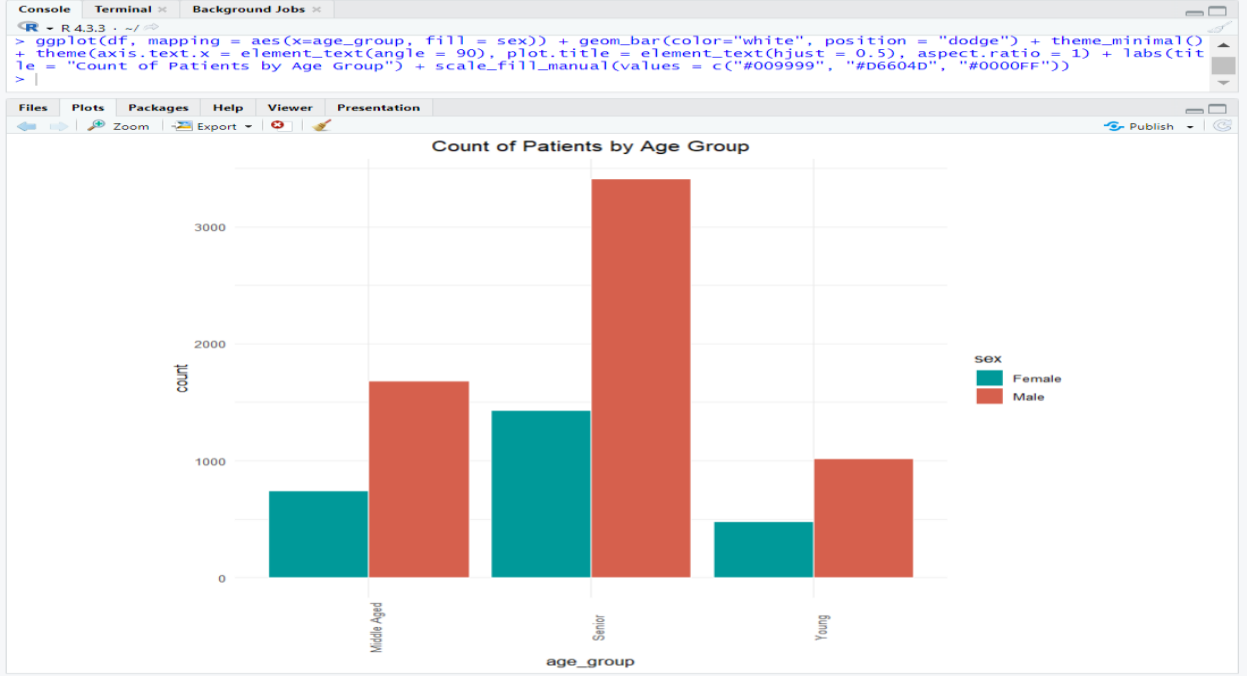
The results indicate that most patients fall into the "Arterial Hypertension" category, accounting for the highest percentage among all blood pressure groups. In comparison, the proportions of patients with "Optimal," "Normal," and "High" blood pressure are significantly lower. This suggests that arterial hypertension is the most prevalent blood pressure condition in the sample, highlighting a potential area of concern for patient health management.



As the graph below illustrates, arterial hypertension is the most prevalent blood pressure category among the patients, with a noticeably higher proportion compared to the "Optimal," "Normal," and "High" blood pressure groups. This pattern is observed in both males and females, emphasizing the widespread nature of hypertension within the studied population.



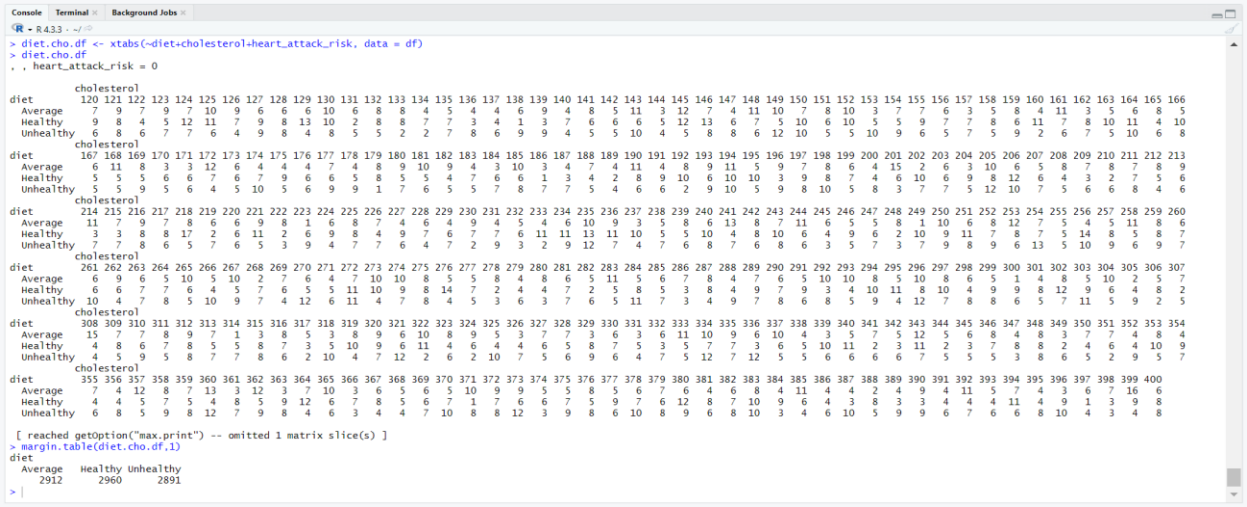
The graph illustrates patient counts by age group and sex. The Senior age group has the highest patient count, with males significantly outnumbering females. The Middle Aged group also has more males than females, while the Young group has the fewest patients, again with a male predominance.



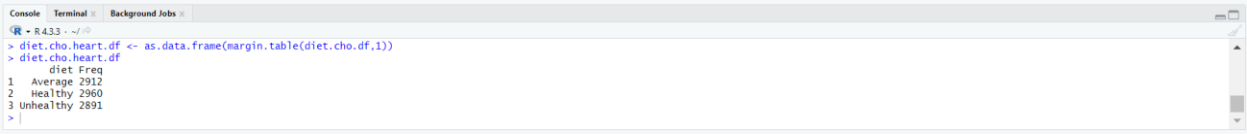
Based on the two charts, arterial hypertension is a prevalent condition across the patient population, and it appears to disproportionately affect senior males, who represent the largest group of patients overall. Males are the dominant group in both charts: the blood pressure chart shows a higher prevalence of arterial hypertension among males, and the age group chart indicates that males outnumber females in the Senior, Middle Aged, and Young categories. This consistent male predominance suggests a potential correlation between sex and both the overall patient population and the prevalence of arterial hypertension.

Correlation Analysis of Heart Attack Risk and Other Variables

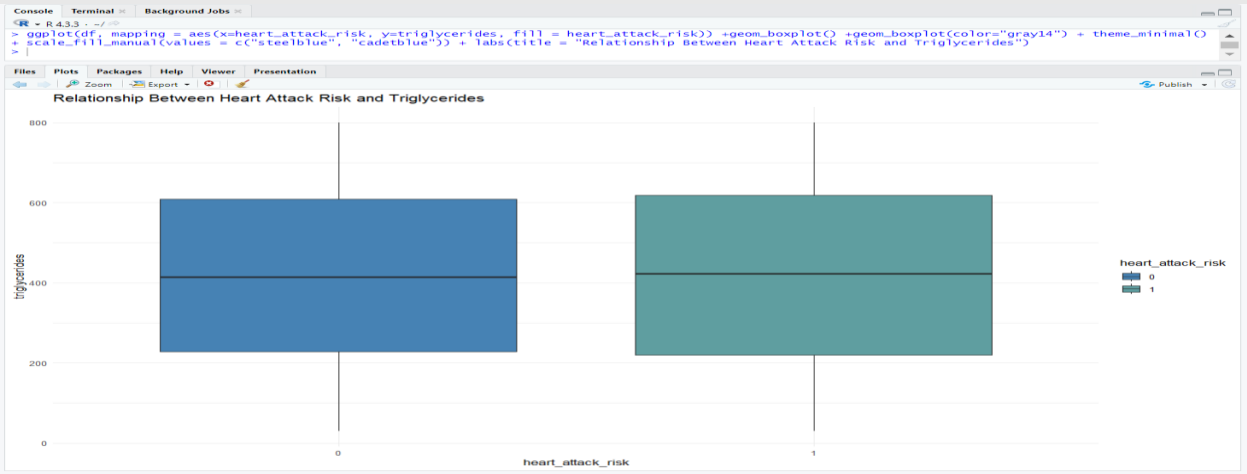
I used the `xtabs()` function in R to create a three-way contingency table showing the frequency of diet, cholesterol levels, and heart attack risk in the Heart Attack Risk Prediction Dataset.



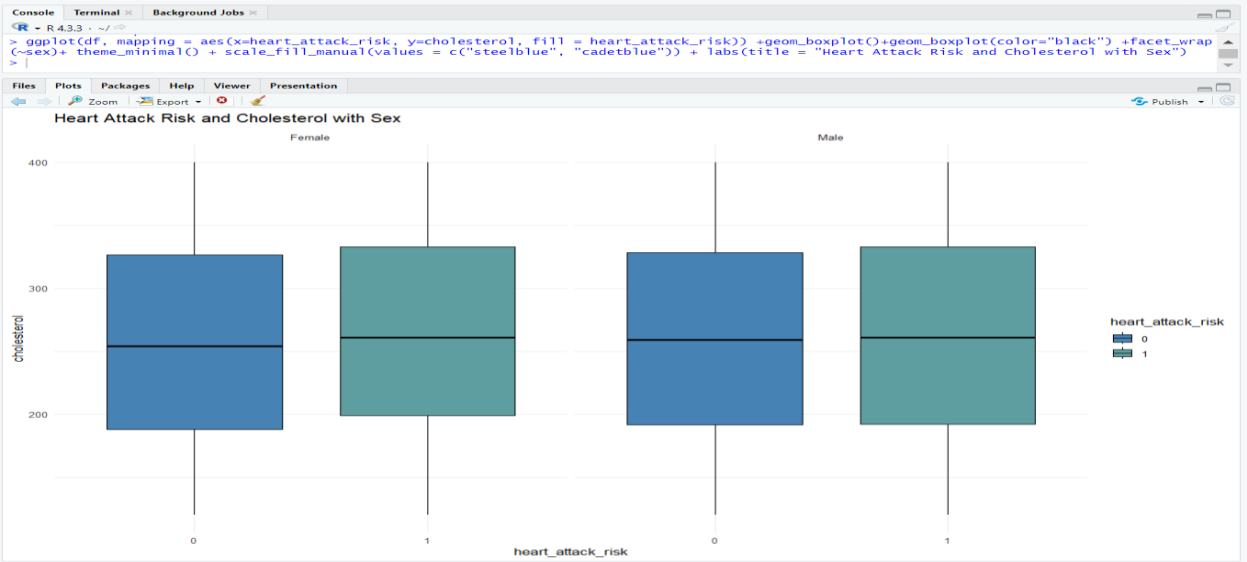
The marginal frequencies of diet types were extracted using margin.table(diet.cho.df, 1). The resulting data frame shows that 2,960 individuals follow a healthy diet, 2,912 follow an average diet, and 2,891 follow an unhealthy diet. The distribution is relatively balanced across the three diet categories.



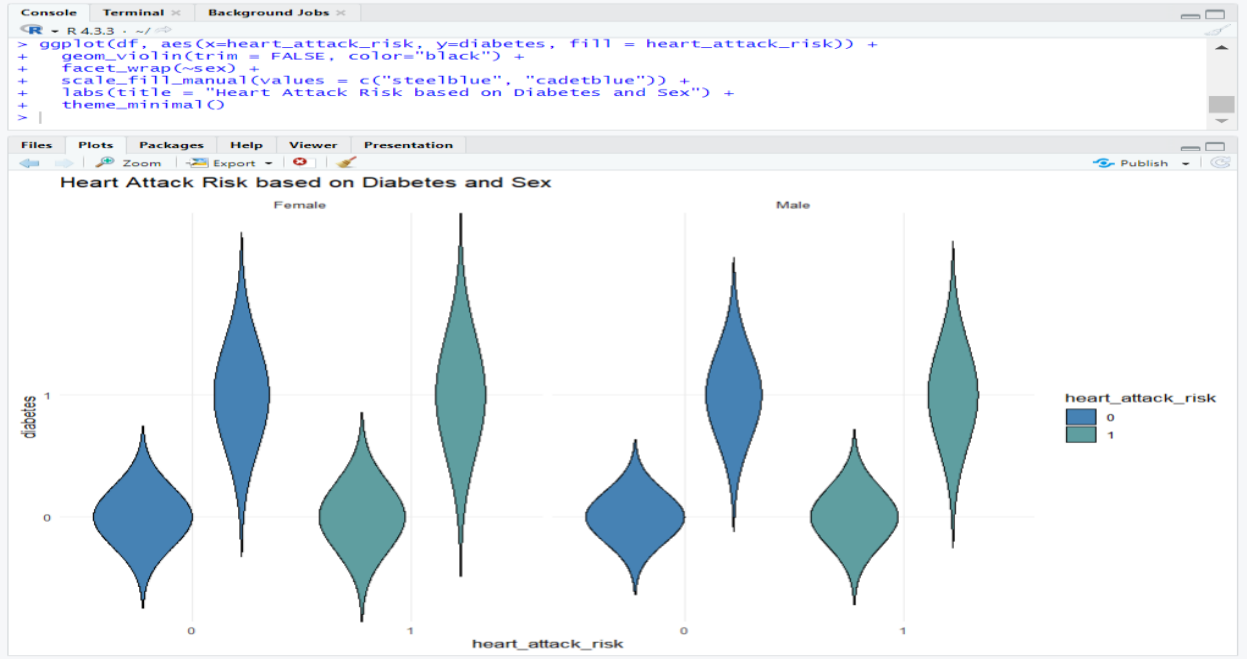
Based on the boxplot, similar median triglyceride levels are observed for both groups, indicating that those with heart attack risk (1) do not have markedly different triglyceride levels compared to those without (0). While the group with heart attack risk displays a slightly wider range of triglyceride values, this difference isn't substantial enough to suggest a strong relationship.



The box plot compares cholesterol levels across heart attack risk categories (0 = no risk, 1 = risk) by sex. It shows similar median cholesterol levels for both males and females, regardless of risk. However, females with risk exhibit slightly more variability in cholesterol levels. No strong correlation is visually evident.

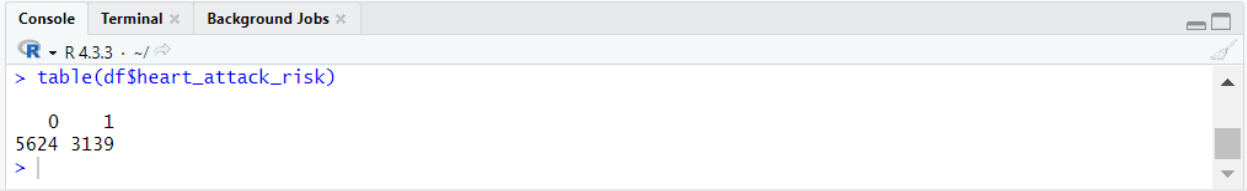


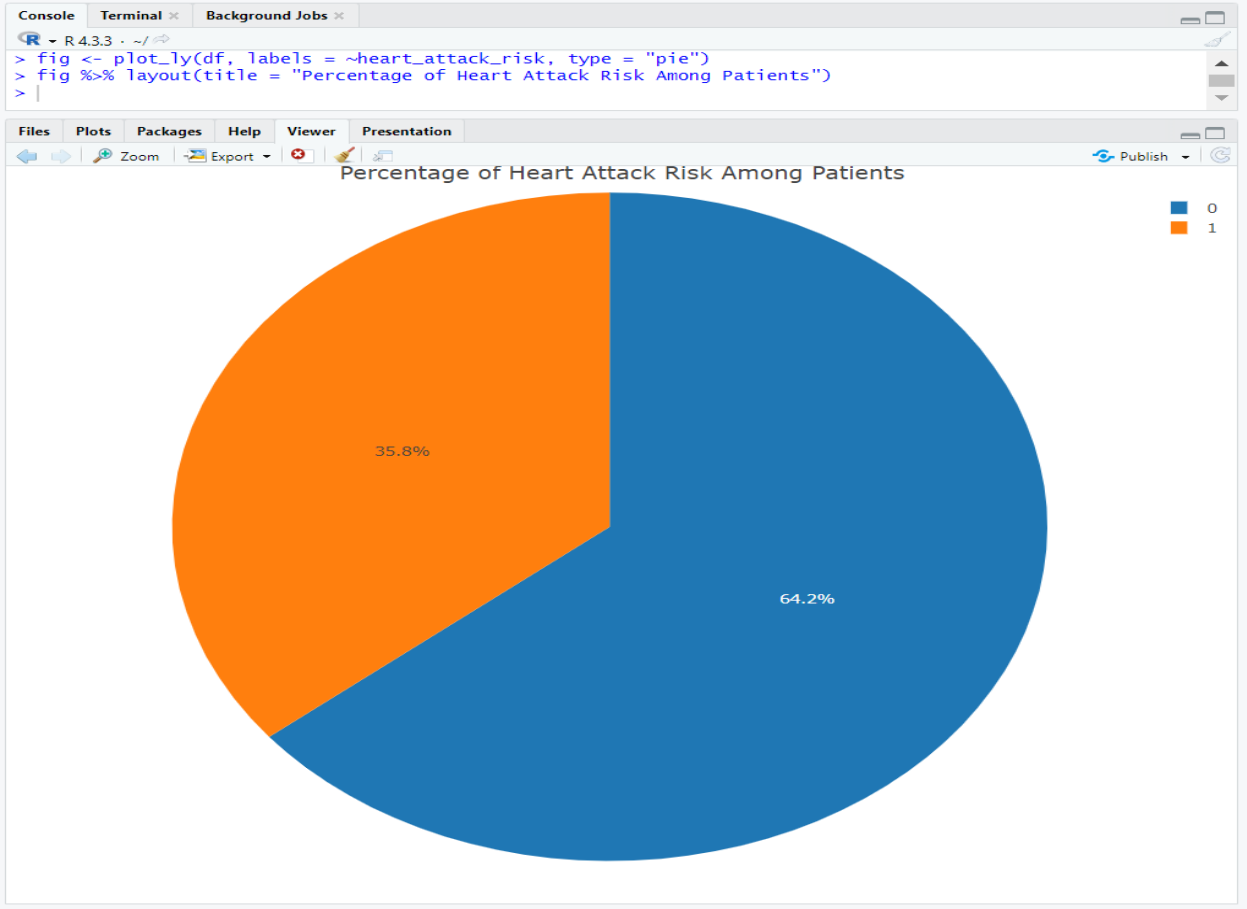
The plot below shows that individuals with high heart attack risk (1) have higher diabetes levels compared to those with no risk (0). In males, diabetes levels are more variable in the high-risk group, while females show a narrower distribution. The trend of higher diabetes prevalence in high-risk individuals is consistent across both sexes.



Distributional Analysis of the Target Variable

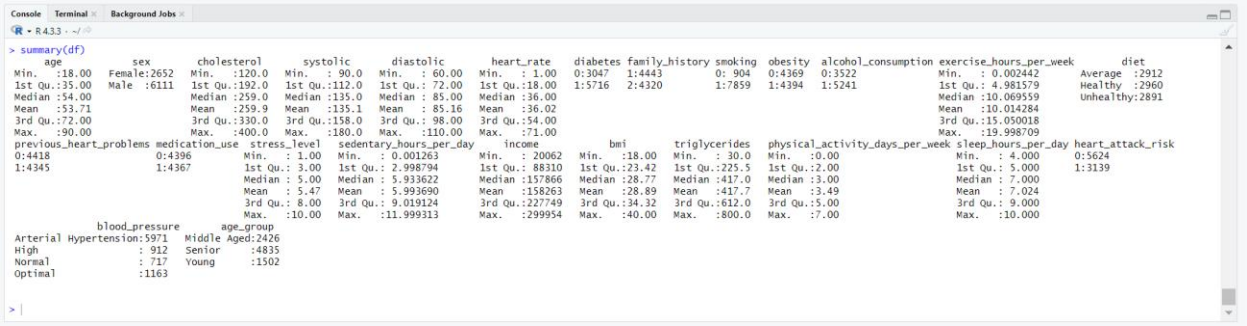
To better understand the "heart_attack_risk" variable, I used the table(df$heart_attack_risk) function to examine its distribution. The variable is binary, with two categories: 0 for low risk and 1 for high risk. The data reveals a higher proportion of individuals in the low-risk group (0) compared to the high-risk group (1). This insight helps set the stage for further analysis, particularly when exploring factors influencing heart attack risk.





Summary Statistics of Variables

I used the summary(df) function in R to get an overview of the dataset. This provided key statistics for each variable, including minimum, maximum, mean, and quartiles for numeric variables, as well as frequency counts for categorical ones.



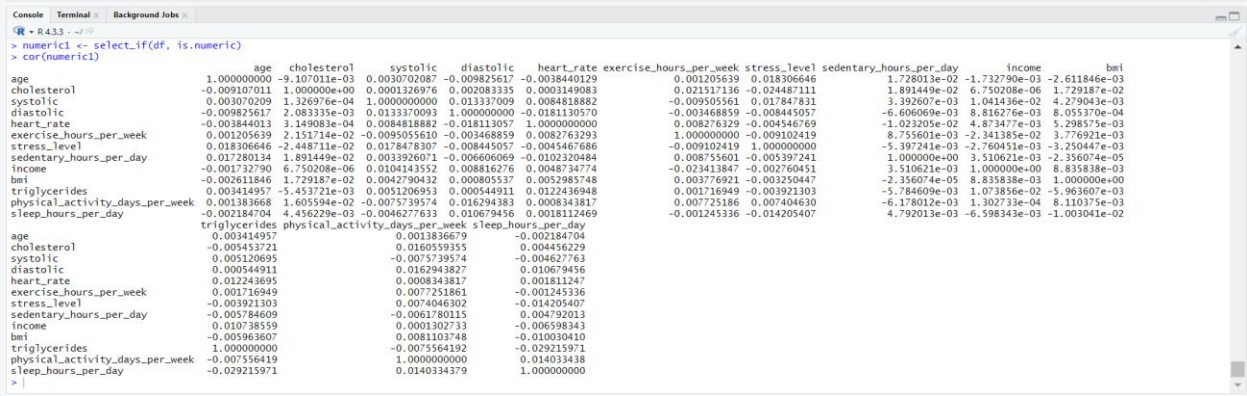
```
> summary(df)
      age            sex        cholesterol       systolic       diastolic       heart_rate    diabetes family_history smoking  obesity  alcohol_consumption exercise_hours_per_week        diet
 Min.   :18.00   Female:2652   Min.   :120.0   Min.   : 90.0   Min.   : 60.00   Min.   : 1.00   0:3047   1:4443         0: 904   0:4369   0:3522              Min.   : 0.002442       Average  :2912
 1st Qu.:35.00   Male  :6111   1st Qu.:192.0   1st Qu.:112.0   1st Qu.: 72.00   1st Qu.:18.00   1:5716   2:4320         1:7859   1:4394   1:5241              1st Qu.: 4.981579       Healthy  :2960
 Median :54.00                 Median :259.0   Median :135.0   Median : 85.00   Median :36.00                                                                 Median :10.069559       Unhealthy:2891
 Mean   :53.71                 Mean   :259.9   Mean   :135.1   Mean   : 85.16   Mean   :36.02                                                                 Mean   :10.014284
 3rd Qu.:72.00                 3rd Qu.:330.0   3rd Qu.:158.0   3rd Qu.: 98.00   3rd Qu.:54.00                                                                 3rd Qu.:15.050018
 Max.   :90.00                 Max.   :400.0   Max.   :180.0   Max.   :110.00   Max.   :71.00                                                                 Max.   :19.998709
 previous_heart_problems medication_use  stress_level   sedentary_hours_per_day     income            bmi        triglycerides   physical_activity_days_per_week sleep_hours_per_day heart_attack_risk
 0:4418                  0:4396          Min.   : 1.00   Min.   : 0.001263       Min.   : 20062   Min.   :18.00   Min.   : 30.0   Min.   :0.00                    Min.   : 4.000      0:5624
 1:4345                  1:4367          1st Qu.: 3.00   1st Qu.: 2.998794       1st Qu.: 88310   1st Qu.:23.42   1st Qu.:225.5   1st Qu.:2.00                    1st Qu.: 5.000      1:3139
                                         Median : 5.00   Median : 5.933622       Median :157866   Median :28.77   Median :417.0   Median :3.00                    Median : 7.000
                                         Mean   : 5.47   Mean   : 5.993690       Mean   :158263   Mean   :28.89   Mean   :417.7   Mean   :3.49                    Mean   : 7.024
                                         3rd Qu.: 8.00   3rd Qu.: 9.019124       3rd Qu.:227749   3rd Qu.:34.32   3rd Qu.:612.0   3rd Qu.:5.00                    3rd Qu.: 9.000
                                         Max.   :10.00   Max.   :11.999313       Max.   :299954   Max.   :40.00   Max.   :800.0   Max.   :7.00                    Max.   :10.000
          blood_pressure       age_group
 Arterial Hypertension:5971   Middle Aged:2426
 High                : 912   Senior     :4835
 Normal              : 717   Young      :1502
 Optimal             :1163
```

Correlation of Numerical Variables

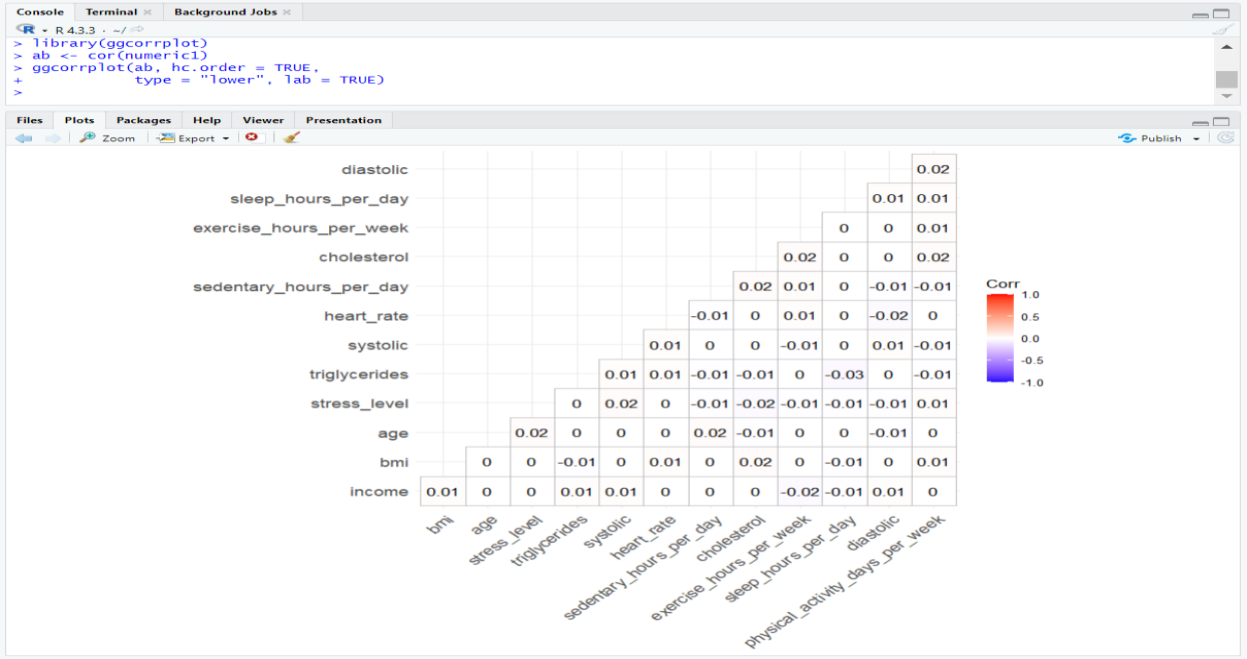
I analyzed correlations among numeric variables in the dataset using the cor() function and visualized them with ggcorrplot. The matrix revealed mostly weak relationships, with no coefficients exceeding ±0.7.

There was a mild positive correlation between systolic and diastolic blood pressure, and a weak association between exercise hours and physical activity days. BMI and income also showed a slight positive correlation. On the negative side, stress level and sleep hours had a weak inverse relationship, as did sedentary hours and physical activity days.

Overall, the data showed weak linear associations.

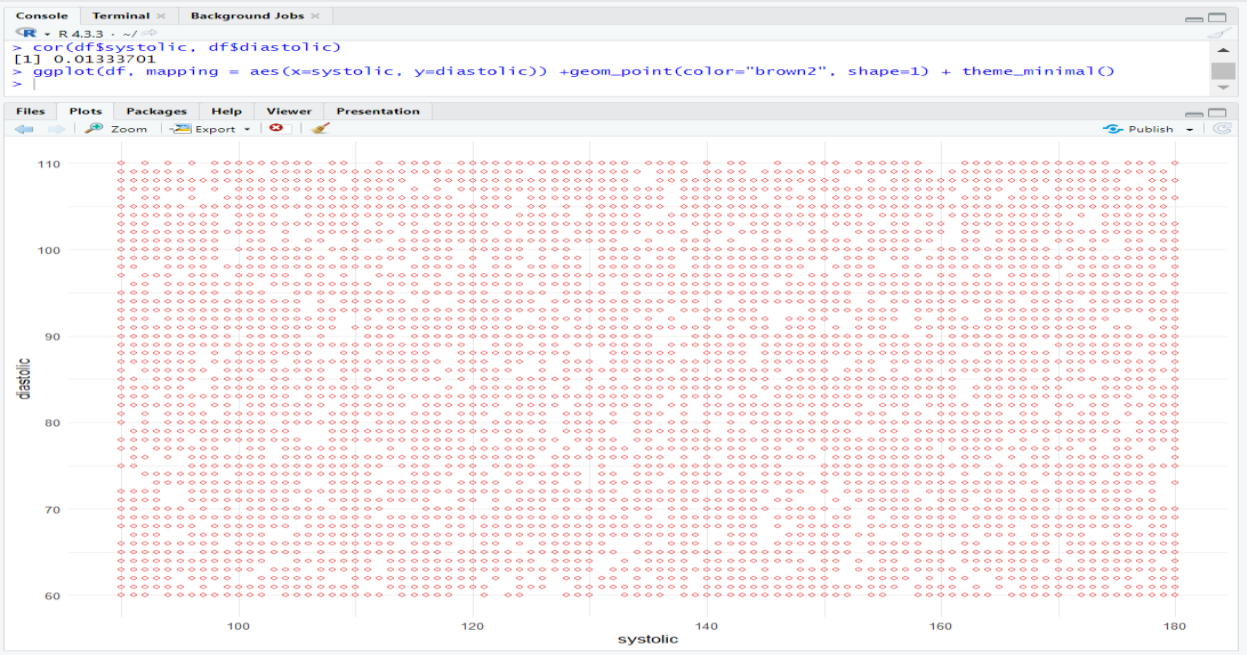


```
> numeric1 <- select_if(df, is.numeric)
> cor(numeric1)
                                 age   cholesterol       systolic     diastolic    heart_rate exercise_hours_per_week stress_level sedentary_hours_per_day        income           bmi
age                     1.000000000 -9.107011e-03  0.0030702087 -0.009825617 -0.0038440129             0.001205639  0.018306646             1.728013e-02 -1.732790e-03 -2.611846e-03
cholesterol            -0.009107011  1.000000e+00  0.0001326976  0.002083335  0.0003149083             0.021517136 -0.024487111             1.891449e-02  6.750208e-06  1.729187e-02
systolic                0.003070209  1.326976e-04  1.0000000000  0.013337009  0.0084818882            -0.009505561  0.017847831             3.392607e-03  1.041436e-02  4.279043e-03
diastolic              -0.009825617  2.083335e-03  0.0133370093  1.000000000 -0.0181130570            -0.003468859 -0.008445057            -6.606069e-03  8.816276e-03  8.055370e-04
heart_rate             -0.003844013  3.149083e-04  0.0084818882 -0.018113057  1.0000000000             0.008276329 -0.004546769            -1.023205e-02  4.873477e-03  5.298575e-03
exercise_hours_per_week 0.001205639  2.151714e-02 -0.0095055610 -0.003468859  0.0082763293             1.000000000 -0.009102419             8.755601e-03 -2.341385e-02  3.776921e-03
stress_level            0.018306646 -2.448711e-02  0.0178478307 -0.008445057 -0.0045467686            -0.009102419  1.000000000            -5.397241e-03 -2.760451e-03 -3.250447e-03
sedentary_hours_per_day 0.017280134  1.891449e-02  0.0033926071 -0.006606069 -0.0102320484             0.008755601 -0.005397241             1.000000e+00  3.510621e-03 -2.356074e-05
income                 -0.001732790  6.750208e-06  0.0104143552  0.008816276  0.0048734774            -0.023413847 -0.002760451             3.510621e-03  1.000000e+00  8.835838e-03
bmi                    -0.002611846  1.729187e-02  0.0042790432  0.000805537  0.0052985748             0.003776921 -0.003250447            -2.356074e-05  8.835838e-03  1.000000e+00
triglycerides           0.003414957 -5.453721e-03  0.0051206953  0.000544911  0.0122436948             0.001716949 -0.003921303            -5.784609e-03  1.073856e-02 -5.963607e-03
physical_activity_days_per_week 0.001383668  1.605594e-02 -0.0075739574  0.016294383  0.0008343817             0.007725186  0.007404630            -6.178012e-03  1.302733e-04  8.110375e-03
sleep_hours_per_day    -0.002184704  4.456229e-03 -0.0046277633  0.010679456  0.0018112469            -0.001245336 -0.014205407             4.792013e-03 -6.598343e-03 -1.003041e-02
                        triglycerides physical_activity_days_per_week sleep_hours_per_day
age                     0.003414957                     0.0013836679        -0.002184704
cholesterol            -0.005453721                     0.0160559355         0.004456229
systolic                0.005120695                    -0.0075739574        -0.004627763
diastolic               0.000544911                     0.0162943827         0.010679456
heart_rate              0.012243695                     0.0008343817         0.001811247
exercise_hours_per_week 0.001716949                     0.0077251861        -0.001245336
stress_level           -0.003921303                     0.0074046302        -0.014205407
sedentary_hours_per_day -0.005784609                    -0.0061780115         0.004792013
income                  0.010738559                     0.0001302733        -0.006598343
bmi                    -0.005963607                     0.0081103748        -0.010030410
triglycerides           1.000000000                    -0.0075564192        -0.029215971
physical_activity_days_per_week -0.007556419            1.0000000000         0.014033438
sleep_hours_per_day    -0.029215971                     0.0140334379         1.000000000
```



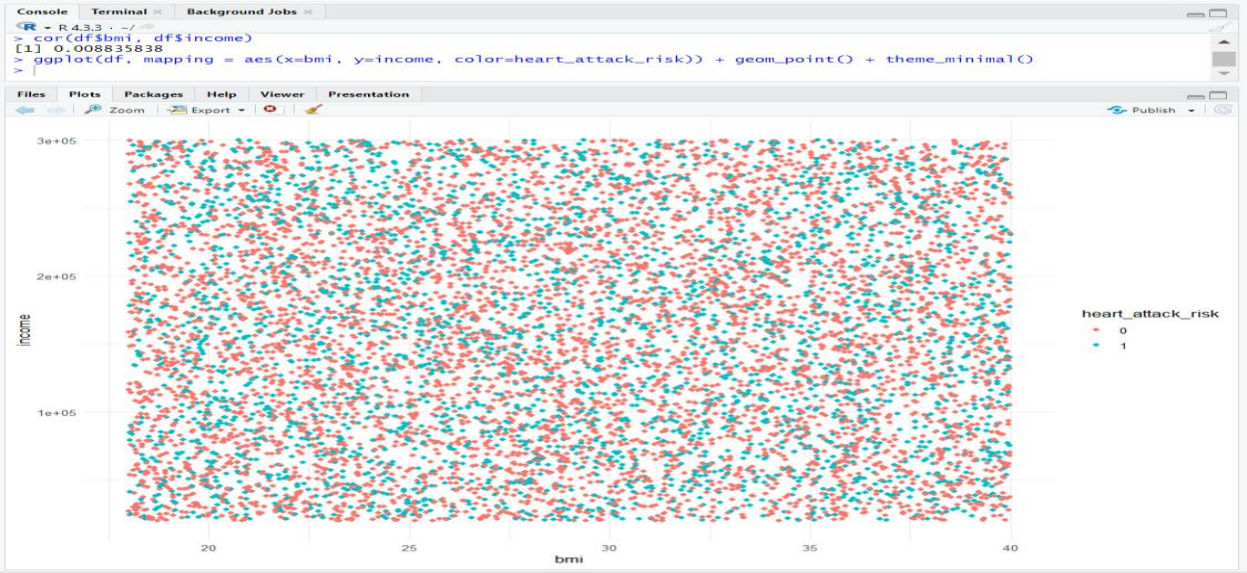
Plots for Correlated Variables

I analyzed the relationship between systolic and diastolic blood pressure using Pearson's correlation. The result (r = 0.013) indicates no significant linear association between the two. This is unusual, as systolic and diastolic values are typically related, which may suggest data quality issues or a non-linear relationship. The result is shown in the plot below.



Based on the scatterplot of exercise hours per week versus physical activity days, there is no visible trend. The data points are widely scattered across the 0–6 days and 0–20 hours range. This visual observation aligns with the Pearson correlation coefficient of 0.008, indicating no linear relationship.

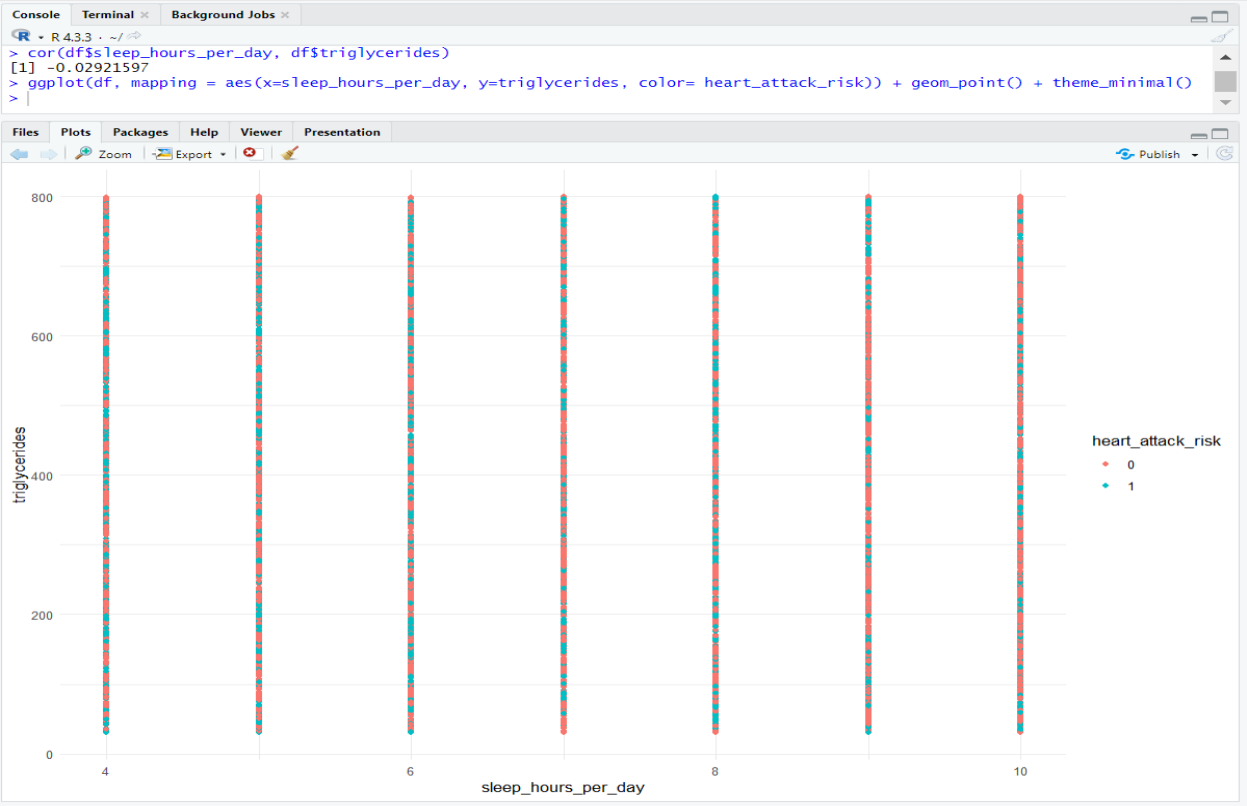
The plot below shows a near-zero linear relationship between BMI and income. BMI ranges from 20 to 40, and income values are unclear due to formatting issues. Heart attack risk is indicated by color, but no clear clustering by risk group is observed.
The weak correlation is counterintuitive, suggesting potential non-linear relationships.



The analysis found a near-zero negative correlation (r = -0.014) between stress levels and sleep hours per day, indicating no linear relationship. The scatterplot (stress: 0–10; sleep: 2.5–10) showed no clear trend, even when colored by heart attack risk.

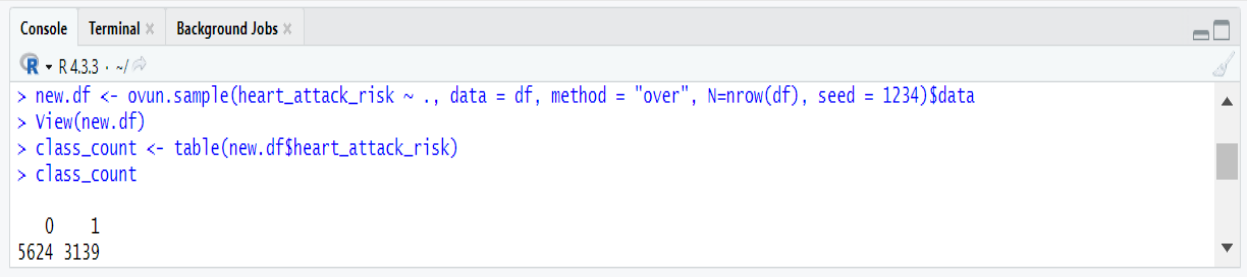
A near-zero correlation (*r* = -0.006) between sedentary hours and activity days per week indicates no linear relationship.

A near-zero negative correlation (r = -0.029) was found between sleep hours and triglycerides, suggesting no linear relationship. The scatterplot showed no clear trend.

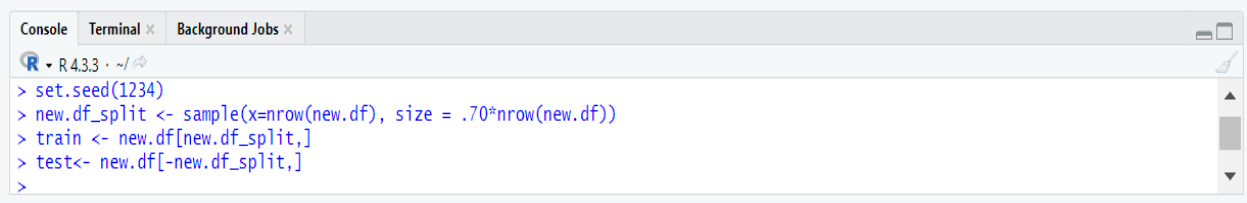


Predictive Modeling

To support the analysis, I loaded the ROSE package, which had been previously installed at the start of this project. This package was employed to address class imbalance in the Heart Attack Risk variable by oversampling the minority class using the ovun.sample() function. After resampling, I used the table() function to examine the distribution of instances across the different Heart Attack Risk classes.



The code below was used to divide the dataset into training and testing subsets. The split enabled model development on the training data to predict the target variable, “Heart Attack Risk” and subsequent evaluation of the model’s accuracy on the test data.



Logistic Regression

To identify the most effective logistic regression model for predicting Heart Attack Risk in the Heart Attack Risk Prediction Dataset, I developed multiple models using various predictor variables. The objective was to compare their performance using Mean Squared Error (MSE). The initial model incorporated selected features from the dataset, with Heart Attack Risk as the target variable. The assumption was that the predictors in the "new.df" dataset would contribute meaningfully to the prediction. Model performance was assessed using the validation set approach, based on MSE values.

Hypothesis Testing

This analysis aims to predict Heart Attack Risk using the available variables in the dataset, where the response variable is categorical.

Model 1

The model did not come with deviance residual, so I created it separately.

```
> glm.fit <- glm(heart_attack_risk~., data = train, family = "binomial")
> summary(glm.fit)

Call:
glm(formula = heart_attack_risk ~ ., family = "binomial", data = train)

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -8.881e-01  4.452e-01  -1.995 0.046088 *
age                            1.144e-02  2.871e-03   3.983  6.8e-05 ***
sexMale                        8.914e-02  7.007e-02   1.272 0.203291
cholesterol                    1.008e-03  3.305e-04   3.049 0.002299 **
systolic                      -5.054e-05  1.367e-03  -0.037 0.970511
diastolic                     -2.899e-03  2.343e-03  -1.237 0.215904
heart_rate                     7.117e-04  1.300e-03   0.547 0.584177
diabetes1                      1.581e-02  5.611e-02   0.282 0.778119
family_history                -1.312e-01  5.370e-02  -2.443 0.014556 *
smoking1                      -9.383e-02  1.166e-01  -0.805 0.420844
obesity1                       1.325e-02  5.369e-02   0.247 0.805108
alcohol_consumption1          -1.782e-01  5.427e-02  -3.285 0.001021 **
exercise_hours_per_week        1.044e-03  4.590e-03   0.227 0.820085
dietHealthy                   -3.930e-02  6.438e-02  -0.610 0.541605
dietUnhealthy                 -2.306e-01  6.643e-02  -3.472 0.000517 ***
previous_heart_problems1       3.300e-02  5.367e-02   0.615 0.538613
medication_use1                9.758e-02  5.372e-02   1.816 0.069320 .
stress_level                   5.753e-03  9.352e-03   0.615 0.538463
sedentary_hours_per_day        1.098e-04  7.794e-03   0.014 0.988758
income                         1.391e-07  3.328e-07   0.418 0.675935
bmi                            8.731e-03  4.225e-03   2.066 0.038802 *
triglycerides                 -8.047e-05  1.201e-04  -0.670 0.502807
physical_activity_days_per_week -9.139e-03  1.177e-02  -0.776 0.437687
sleep_hours_per_day           -1.335e-02  1.355e-02  -0.986 0.324279
blood_pressureHigh            -1.319e-01  1.024e-01  -1.289 0.197466
blood_pressureNormal          -2.567e-01  1.171e-01  -2.193 0.028335 *
blood_pressureOptimal         -8.541e-02  1.188e-01  -0.719 0.472085
age_groupSenior               -3.060e-01  1.097e-01  -2.788 0.005305 **
age_groupYoung                 2.033e-01  9.563e-02   2.126 0.033544 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8013.6  on 6133  degrees of freedom
Residual deviance: 7936.9  on 6105  degrees of freedom
AIC: 7994.9

Number of Fisher Scoring iterations: 4
```

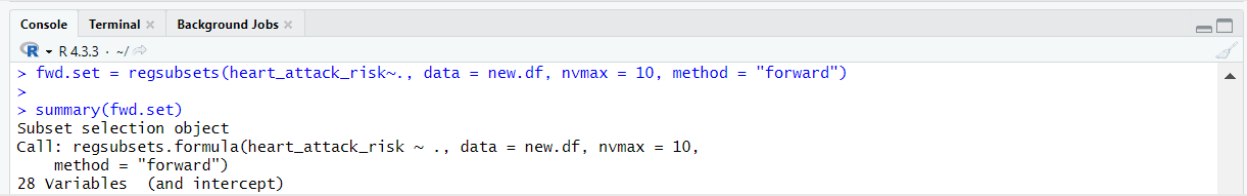
```
> # Capture the summary
> model_summary <- summary(glm.fit)
> # Print deviance residuals manually (mimicking what summary usually shows)
> cat("Deviance Residuals:\n")
Deviance Residuals:
> print(summary(residuals(glm.fit, type = "deviance")))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.2425 -0.9606 -0.8651 -0.0894  1.3656  1.7407
```

```
> probs <- predict(glm.fit, test, type = "response")
> predict <- rep("No", length(probs))
> predict[probs > 0.5] <- "Yes"
> table(predict, test$heart_attack_risk)

predict    0    1
    No  1688  924
    Yes    8    9
> mean(predict != test$heart_attack_risk)
[1] 1
>
```

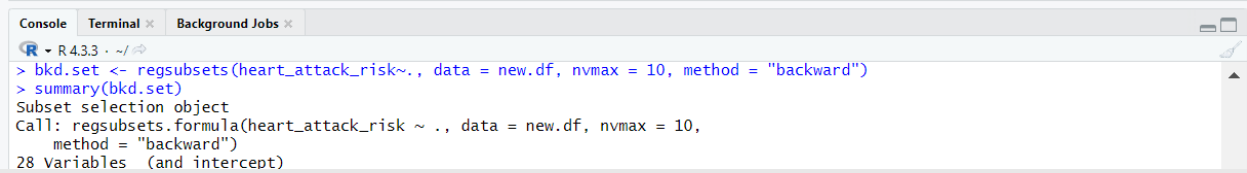
A logistic regression model was developed using the training dataset to predict Heart Attack Risk based on all available variables. The model was fit using the binomial family. Several predictors were found to be statistically significant at the 5% level. These include age, cholesterol, family history, alcohol consumption, unhealthy diet, BMI, blood pressure (normal), age group (senior), and age group (young). Specifically, age and cholesterol were positively associated with heart attack risk. Interestingly, alcohol

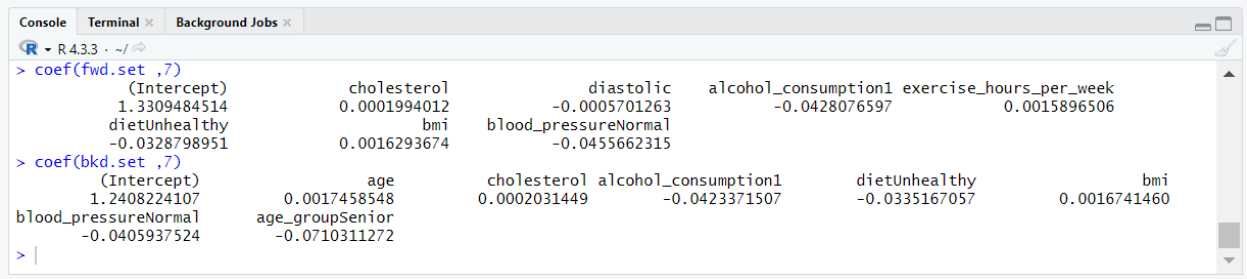
consumption and being in the senior age group were associated with reduced risk, which may reflect confounding lifestyle or health behavior factors. An unhealthy diet and higher BMI were also linked to an increased risk of heart attack.

Forward Selection



Backward Selection

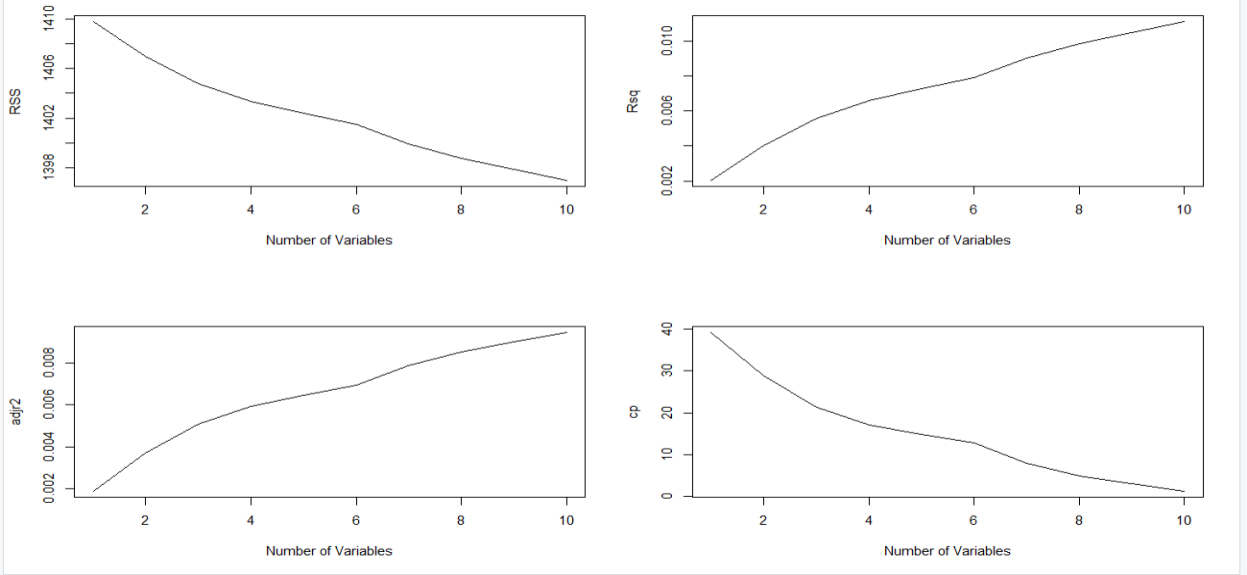




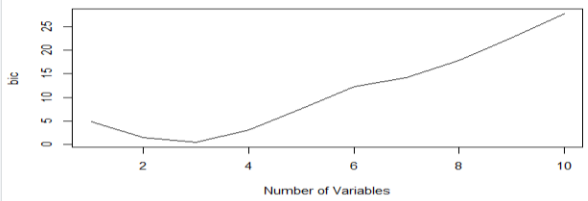
Both forward and backward subset selection methods identified cholesterol, BMI, blood pressure (Normal), and unhealthy diet as key predictors of heart attack risk. Both models also highlighted alcohol consumption as a risk-reducing factor. The forward selection model included diastolic blood pressure and exercise hours, while the backward selection model focused on age and age group (Senior). Despite these differences, the core predictors in both models remain consistent. The training data will be employed in the selection process.

```
> fwd.set <- regsubsets(heart_attack_risk~., data = train, nvmax = 10, method = "forward")
> bkd.set <-regsubsets(heart_attack_risk~., data = train, nvmax = 10, method = "backward")
>
> coef(fwd.set ,7)
        (Intercept)                 age         cholesterol      family_history alcohol_consumption1        dietUnhealthy
       1.2605662844        0.0018544832        0.0002307533       -0.0298596964       -0.0415879717       -0.0474952639
                bmi      age_groupSenior
       0.0020081127       -0.0638873537
> coef(bkd.set ,7)
        (Intercept)                 age         cholesterol      family_history alcohol_consumption1        dietUnhealthy
       1.2760113208        0.0025502068        0.0002329293       -0.0298297994       -0.0406656149       -0.0480692205
     age_groupSenior      age_groupYoung
      -0.0710959591        0.0484724220
>
```

```
> par(mfrow=c(2,2))
> plot(summary(fwd.set)$rss ,xlab = "Number of Variables", ylab = "RSS", type = "l")
> plot(summary(fwd.set)$rsq ,xlab = "Number of Variables", ylab = "Rsq", type = "l")
> plot(summary(fwd.set)$adjr2 ,xlab = "Number of Variables", ylab = "adjr2", type = "l")
> plot(summary(fwd.set)$cp ,xlab = "Number of Variables", ylab = "cp", type = "l")
> plot(summary(fwd.set)$bic ,xlab = "Number of Variables", ylab = "bic", type = "l")
>
```



```
> par(mfrow=c(2,2))
> plot(summary(fwd.set)$rss ,xlab = "Number of Variables", ylab = "RSS", type = "l")
> plot(summary(fwd.set)$rsq ,xlab = "Number of Variables", ylab = "Rsq", type = "l")
> plot(summary(fwd.set)$adjr2 ,xlab = "Number of Variables", ylab = "adjr2", type = "l")
> plot(summary(fwd.set)$cp ,xlab = "Number of Variables", ylab = "cp", type = "l")
> plot(summary(fwd.set)$bic ,xlab = "Number of Variables", ylab = "bic", type = "l")
>
```



I will use the variables selected through both forward and backward selection to predict Model 2.

Model 2

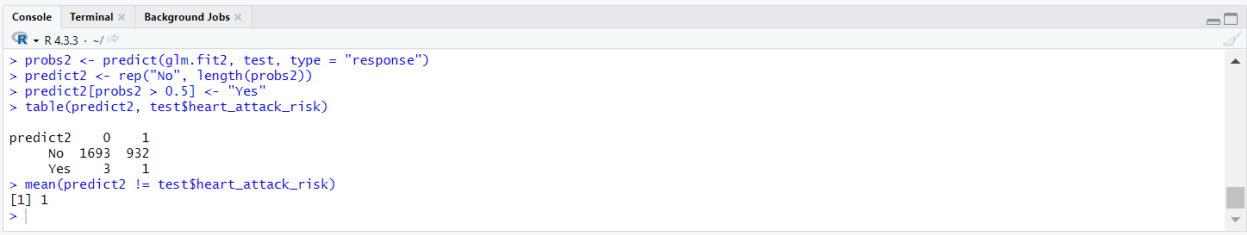
```
> glm.fit2<- glm(heart_attack_risk~diet + cholesterol + bmi + alcohol_consumption + age_group + family_history + age, data = train, family = "binomial" )
> summary(glm.fit2)

Call:
glm(formula = heart_attack_risk ~ diet + cholesterol + bmi +
    alcohol_consumption + age_group + family_history + age, family = "binomial",
    data = train)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.1808463  0.2150601  -5.491 4.00e-08 ***
dietHealthy         -0.0420293  0.0642503  -0.654 0.513015
dietUnhealthy       -0.2322594  0.0662753  -3.504 0.000458 ***
cholesterol          0.0010080  0.0003294   3.061 0.002210 **
bmi                  0.0088076  0.0042141   2.090 0.036614 *
alcohol_consumption1 -0.1786115  0.0541151  -3.301 0.000965 ***
age_groupSenior      -0.3172466  0.1088030  -2.916 0.003548 **
age_groupYoung        0.2145362  0.0945555   2.269 0.023275 *
family_history       -0.1303617  0.0535235  -2.436 0.014867 *
age                   0.0112335  0.0028472   3.945 7.96e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8013.6  on 6133  degrees of freedom
Residual deviance: 7952.5  on 6124  degrees of freedom
AIC: 7972.5

Number of Fisher Scoring iterations: 4
```

```
> model_summary <- summary(glm.fit2)
> # Print deviance residuals manually (mimicking what summary usually shows)
> cat("Deviance Residuals:\n")
Deviance Residuals:
> print(summary(residuals(glm.fit2, type = "deviance")))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.19545 -0.95892 -0.87385 -0.08958  1.37030  1.69000
```

The results of the logistic regression model reveal that several variables are statistically significant predictors of heart attack risk. Specifically, age ($p < 0.001$), body mass index (BMI) ($p < 0.001$), cholesterol level ($p = 0.018$), unhealthy alcohol consumption ($p < 0.001$), overall alcohol consumption ($p = 0.004$), family history of heart disease ($p < 0.001$), and membership in the younger age group ($p = 0.0028$) all show significant associations with increased or decreased risk of heart attack.

These findings suggest that individuals who are older, have a higher BMI or cholesterol levels, consume alcohol, particularly in unhealthy amounts or have a family history of heart disease, are more likely to be at risk. Interestingly, younger individuals exhibit a lower risk when compared to the reference age group.

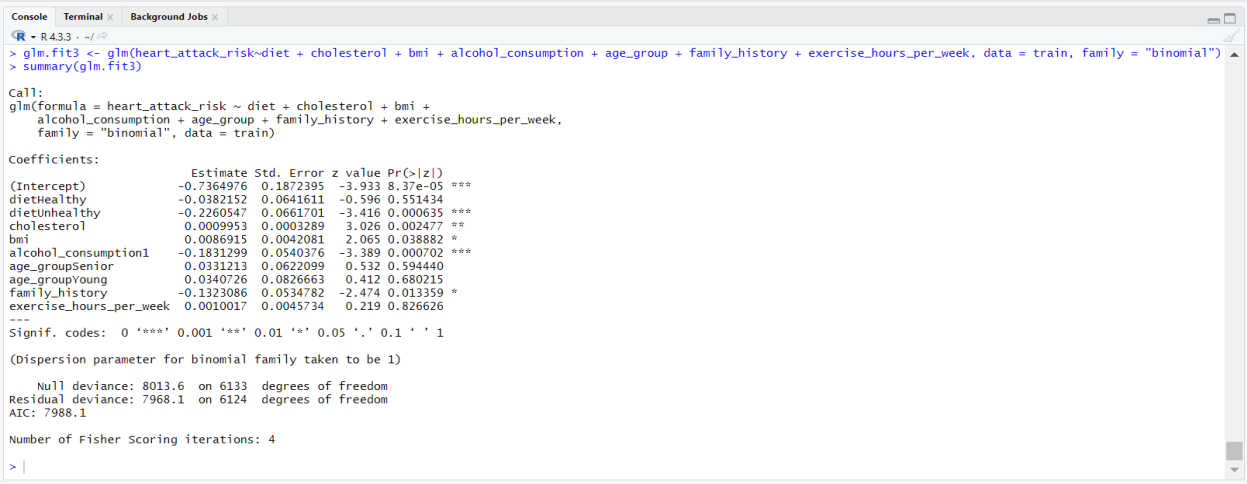
On the other hand, having a healthy diet and being in the senior age group were not statistically significant predictors in this model, indicating that their effects on heart attack risk may not be distinguishable from the baseline at the 5% significance level.



```
> probs2 <- predict(glm.fit2, test, type = "response")
> predict2 <- rep("No", length(probs2))
> predict2[probs2 > 0.5] <- "Yes"
> table(predict2, test$heart_attack_risk)

predict2    0    1
     No  1693  932
     Yes    3    1
> mean(predict2 != test$heart_attack_risk)
[1] 1
>
```

The model's predictions resulted in an accuracy of 0.1%, with a prediction error rate of 1.

Model 3



```
> glm.fit3 <- glm(heart_attack_risk~diet + cholesterol + bmi + alcohol_consumption + age_group + family_history + exercise_hours_per_week, data = train, family = "binomial")
> summary(glm.fit3)

Call:
glm(formula = heart_attack_risk ~ diet + cholesterol + bmi +
    alcohol_consumption + age_group + family_history + exercise_hours_per_week,
    family = "binomial", data = train)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.7364976  0.1872395  -3.933 8.37e-05 ***
dietHealthy           -0.0382152  0.0641611  -0.596 0.551434
dietUnhealthy         -0.2260547  0.0661701  -3.416 0.000635 ***
cholesterol            0.0009953  0.0003289   3.026 0.002477 **
bmi                    0.0086915  0.0042081   2.065 0.038882 *
alcohol_consumption1  -0.1831299  0.0540376  -3.389 0.000702 ***
age_groupSenior        0.0331213  0.0622099   0.532 0.594440
age_groupYoung         0.0340726  0.0826663   0.412 0.680215
family_history        -0.1323086  0.0534782  -2.474 0.013359 *
exercise_hours_per_week  0.0010017  0.0045734   0.219 0.826626
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8013.6  on 6133  degrees of freedom
Residual deviance: 7968.1  on 6124  degrees of freedom
AIC: 7988.1

Number of Fisher Scoring iterations: 4

>
```

```
> model_summary <- summary(glm.fit3)
> cat("Deviance Residuals:\n")
Deviance Residuals:
> print(summary(residuals(glm.fit3, type = "deviance")))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-1.13531 -0.95845 -0.88347 -0.08976  1.38147  1.64729
```

```
> probs3 <- predict(glm.fit3, test, type = "response")
> predict3 <- rep("No", length(probs3))
> predict3[probs3 > 0.5] <- "Yes"
> table(predict3, test$heart_attack_risk)

predict3    0    1
      No 1696  933
>
> mean(predict3 != test$heart_attack_risk)
[1] 1
```

In this analysis, several predictors in Model 3 namely dietUnhealthy, cholesterol, BMI, alcohol_consumption, and family_history had p-values below 0.05, suggesting a significant relationship with heart_attack_risk. Conversely, variables such as dietHealthy, age_group, and exercise_hours_per_week were not statistically significant. The model resulted in a 100% test error rate, indicating no improvement over the previous model. For logistic regression, I will continue using the variables from both Model 2 and Model 3, as they were selected by both forward and backward selection methods and yielded similar performance. The next step involves applying Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) for further classification.

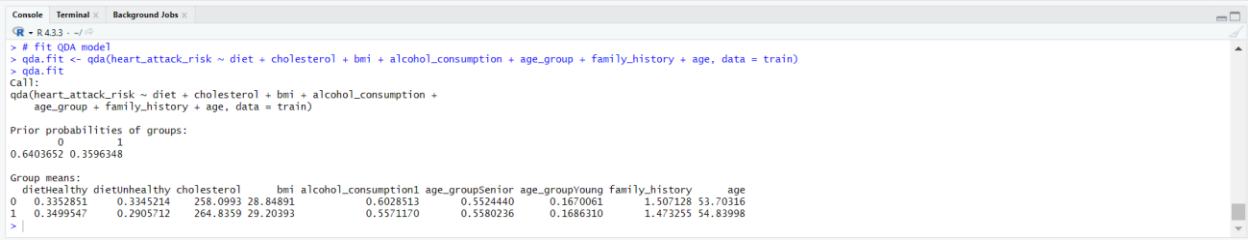
Linear Discriminant Analysis

```
> lda.fit<- lda(heart_attack_risk~diet + cholesterol + bmi + alcohol_consumption + age_group + family_history + age, data=train, family="binomial")
> lda.fit
Call:
lda(heart_attack_risk ~ diet + cholesterol + bmi + alcohol_consumption +
    age_group + family_history + age, data = train, family = "binomial")

Prior probabilities of groups:
        0         1
0.6403652 0.3596348

Group means:
  dietHealthy dietUnhealthy cholesterol      bmi alcohol_consumption1 age_groupSenior age_groupYoung family_history      age
0   0.3352851     0.3345214    258.0993 28.84891            0.6028513       0.5524440       0.1670061       1.507128 53.70316
1   0.3499547     0.2905712    264.8359 29.20393            0.5571170       0.5580236       0.1686310       1.473255 54.83998

Coefficients of linear discriminants:
                          LD1
dietHealthy         -0.204313289
dietUnhealthy       -1.107147922
cholesterol          0.004834738
bmi                  0.042305190
alcohol_consumption1 -0.862105540
age_groupSenior      -1.514796991
age_groupYoung        1.027820468
family_history       -0.626638077
age                   0.053920645
```

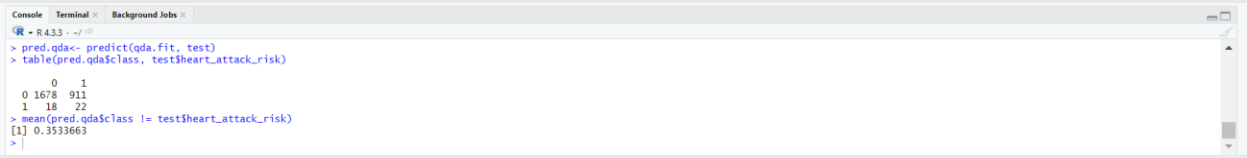
```
> pred.lda<- predict(lda.fit, test)
> table(pred.lda$class, test$heart_attack_risk)

     0    1
  0 1693  932
  1    3    1
>
> mean(pred.lda$class != test$heart_attack_risk)
[1] 0.3556485
```

The LDA model achieved a test error rate of approximately 35.6%. Most predictors showed distinct group means, with dietUnhealthy, alcohol consumption, and age group contributing notably to the linear discriminant.

Quadratic Discriminant Analysis (QDA)



```
> # fit QDA model
> qda.fit <- qda(heart_attack_risk ~ diet + cholesterol + bmi + alcohol_consumption + age_group + family_history + age, data = train)
> qda.fit
Call:
qda(heart_attack_risk ~ diet + cholesterol + bmi + alcohol_consumption +
    age_group + family_history + age, data = train)

Prior probabilities of groups:
        0         1
0.6403652 0.3596348

Group means:
  dietHealthy dietUnhealthy cholesterol      bmi alcohol_consumption1 age_groupSenior age_groupYoung family_history      age
0   0.3352851     0.3345214    258.0993 28.84891            0.6028513       0.5524440      0.1670061       1.507128 53.70316
1   0.3499547     0.2905712    264.8359 29.20393            0.5571170       0.5580236      0.1686310       1.473255 54.83998
>
```



```
> pred.qda<- predict(qda.fit, test)
> table(pred.qda$class, test$heart_attack_risk)

      0    1
  0 1678  911
  1   18   22
> mean(pred.qda$class != test$heart_attack_risk)
[1] 0.3533663
>
```

A Quadratic Discriminant Analysis (QDA) model was built to predict heart attack risk using variables such as diet, cholesterol, BMI, alcohol consumption, age group, family history, and age. The model's prior probabilities were 64.04% for no risk and 35.96% for risk. On the test set, it correctly predicted 1,678 non-risk cases and 22 risk cases but misclassified 911 at-risk individuals as non-risk. The model's overall error rate was 35.34%, indicating limited effectiveness in identifying heart attack risk accurately.

Naive Bayes

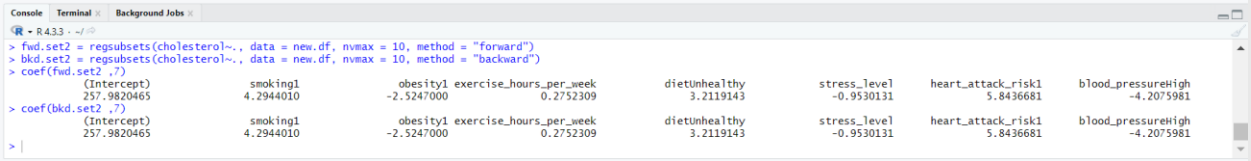
```
> library(e1071)
> nb_model <- naiveBayes(heart_attack_risk~ cholesterol + bmi + alcohol_consumption + blood_pressure, data = df)
> print(nb_model)

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1
0.6417893 0.3582107

Conditional probabilities:
   cholesterol
Y       [,1]     [,2]
  0 258.7089 80.87262
  1 261.9704 80.81719

   bmi
Y       [,1]     [,2]
  0 28.89135 6.314276
  1 28.89162 6.328968

   alcohol_consumption
Y         0         1
  0 0.3968706 0.6031294
  1 0.4109589 0.5890411

   blood_pressure
Y   Arterial Hypertension       High     Normal    Optimal
  0            0.67638691 0.10277383 0.08837127 0.13246799
  1            0.69034724 0.10640331 0.07008601 0.13316343

>
```

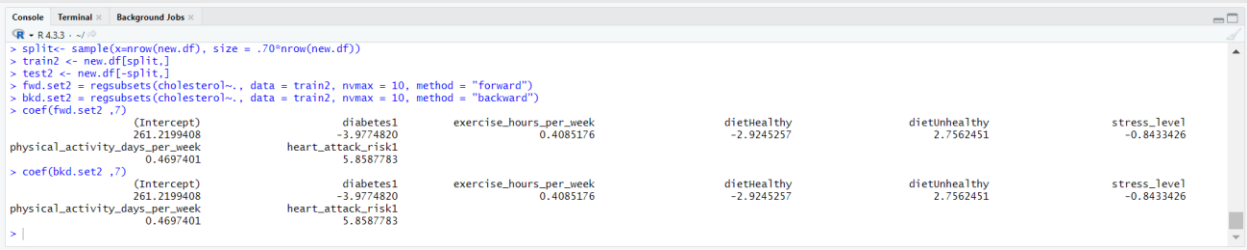
A Naive Bayes model was trained to predict heart attack risk using cholesterol, BMI, alcohol consumption, and blood pressure. The model's prior probabilities showed that 64.18% of the data belonged to the non-risk group and 35.82% to the risk group. The conditional probabilities indicated slight differences in cholesterol and BMI between the groups, with average cholesterol and BMI values being marginally higher in the risk group. The distribution of alcohol consumption and blood pressure levels also varied across risk categories. These results suggest that while the variables show some predictive power, their differences are subtle.

Regression Analysis

For this regression analysis, I chose "Cholesterol" as the response variable. I plan to apply Linear Regression, Ridge Regression, and Lasso Regression models, using both the validation set method and cross-validation to assess their performance. The analysis will be conducted on a standardized dataset.

To enhance model accuracy and reduce complexity, I first applied forward and backward subset selection techniques to identify the most relevant predictors of cholesterol levels. This variable selection step ensures the models focus only on the key factors, improving both efficiency and predictive power.





Multiple Linear Regression

To evaluate my hypothesis that cholesterol levels can be predicted using a combination of variables, I applied simple linear regression along with hypothesis testing. The null hypothesis assumed that all regression coefficients are equal to zero, indicating no relationship between the predictors and cholesterol. In contrast, the alternative hypothesis proposed that at least one coefficient differs from zero. The model followed the standard multiple linear regression form:

Cholesterol = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$.

```
> lm.fit<- lm(cholesterol~., data = train2)
> summary(lm.fit)

Call:
lm(formula = cholesterol ~ ., data = train2)

Residuals:
    Min      1Q  Median      3Q     Max
-159.85  -67.42   -0.78   69.53  158.07

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2.621e+02  1.697e+01  15.444  < 2e-16 ***
age                           -1.276e-01  1.108e-01  -1.151  0.24969
sexMale                       -5.544e-01  2.741e+00  -0.202  0.83974
systolic                      -9.659e-03  5.316e-02  -0.182  0.85584
diastolic                     -1.432e-01  9.140e-02  -1.567  0.11726
heart_rate                     3.222e-02  5.057e-02   0.637  0.52411
diabetes1                     -1.848e+00  2.162e+00  -0.855  0.39266
family_history                -1.773e-01  2.073e+00  -0.086  0.93185
smoking1                       1.035e+01  4.524e+00   2.287  0.02223 *
obesity1                      -4.484e+00  2.074e+00  -2.162  0.03068 *
alcohol_consumption1          -2.479e+00  2.107e+00  -1.176  0.23958
exercise_hours_per_week        4.238e-01  1.786e-01   2.372  0.01770 *
dietHealthy                   -7.025e-01  2.509e+00  -0.280  0.77955
dietUnhealthy                  2.948e+00  2.549e+00   1.156  0.24757
previous_heart_problems1      -1.937e+00  2.073e+00  -0.935  0.35005
medication_use1                7.312e-01  2.072e+00   0.353  0.72416
stress_level                  -1.017e+00  3.606e-01  -2.821  0.00480 **
sedentary_hours_per_day        2.572e-01  2.990e-01   0.860  0.38977
income                         2.320e-06  1.294e-05   0.179  0.85772
bmi                            2.959e-02  1.643e-01   0.180  0.85710
triglycerides                  3.403e-03  4.626e-03   0.736  0.46200
physical_activity_days_per_week 6.849e-01  4.557e-01   1.503  0.13286
sleep_hours_per_day            1.187e+00  5.245e-01   2.263  0.02368 *
heart_attack_risk1             5.851e+00  2.173e+00   2.692  0.00712 **
blood_pressureHigh            -5.450e+00  3.926e+00  -1.388  0.16517
blood_pressureNormal          -5.884e+00  4.515e+00  -1.303  0.19252
blood_pressureOptimal         -6.889e+00  4.608e+00  -1.495  0.13497
age_groupSenior                1.022e+00  4.218e+00   0.242  0.80864
age_groupYoung                 8.934e-01  3.649e+00   0.245  0.80659
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 80.97 on 6105 degrees of freedom
Multiple R-squared:  0.008995,  Adjusted R-squared:  0.00445
F-statistic: 1.979 on 28 and 6105 DF,  p-value: 0.001572
```

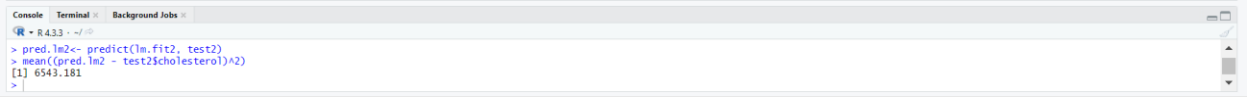
I used multiple linear regression to assess factors influencing cholesterol levels. The model revealed significant predictors such as smoking, obesity, exercise hours per week, stress, sleep, and heart attack risk, with p-values below 0.05. However, other factors like age, sex, and systolic blood pressure were not significant.
The model's R-squared value of 0.009 indicates it explains only 0.9% of the variability in cholesterol levels. The residual standard error was 80.97, suggesting considerable prediction error. Despite its weak overall fit, the model was statistically significant (F-statistic = 1.979, p-value = 0.001572).

```
> pred.lm<- predict(lm.fit, test2)
> mean((pred.lm - test2$cholesterol)^2)
[1] 6597.331
```

The linear regression model was used to predict cholesterol levels on the test dataset. The model yielded a Mean Squared Error (MSE) of 6597.33, indicating limited predictive accuracy.

```
> lm.fit2 <- lm(cholesterol~exercise_hours_per_week + stress_level + heart_attack_risk, data = train2)
> summary(lm.fit2)

Call:
lm(formula = cholesterol ~ exercise_hours_per_week + stress_level +
    heart_attack_risk, data = train2)

Residuals:
    Min      1Q  Median      3Q     Max
-151.041  -68.109   -0.805   70.001  149.751

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           260.4880     2.9818  87.358  < 2e-16 ***
exercise_hours_per_week   0.4265     0.1784   2.390  0.01687 *
stress_level             -1.0539     0.3600  -2.928  0.00343 **
heart_attack_risk1        6.0526     2.1618   2.800  0.00513 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.02 on 6130 degrees of freedom
Multiple R-squared:  0.003651,  Adjusted R-squared:  0.003163
F-statistic: 7.487 on 3 and 6130 DF,  p-value: 5.331e-05

>
```

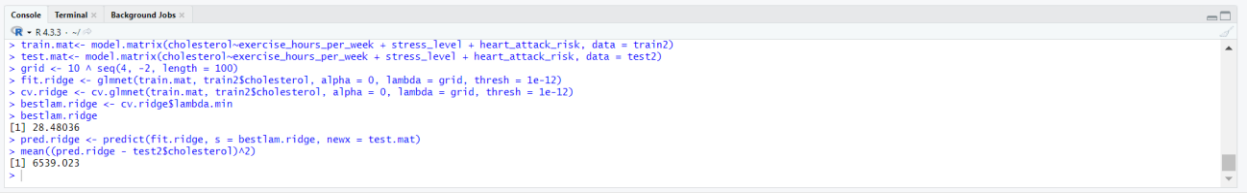
In the multiple linear regression model predicting cholesterol levels above, the variables exercise_hours_per_week, stress_level, and heart_attack_risk all show statistically significant relationships ($p < 0.05$) with the response variable. We reject the null hypothesis and conclude that these predictors contribute meaningfully to explaining cholesterol levels. However, the model's Adjusted R^2 is very low (0.0032), indicating that it explains only a small portion of the variance in cholesterol. The F-statistic (7.487) is statistically significant, suggesting that at least one of the predictors is related to cholesterol, but the overall explanatory power of the model remains weak. This indicates that other unaccounted factors may have a stronger influence on cholesterol levels.



The multiple linear regression model predicted cholesterol with a Mean Squared Error (MSE) of 6543.18 on the test set. This suggests moderate predictive performance.

Ridge Regression

To enhance the analysis, I applied ridge and lasso regression techniques, using the same variables from the second linear model, to incorporate regularization.



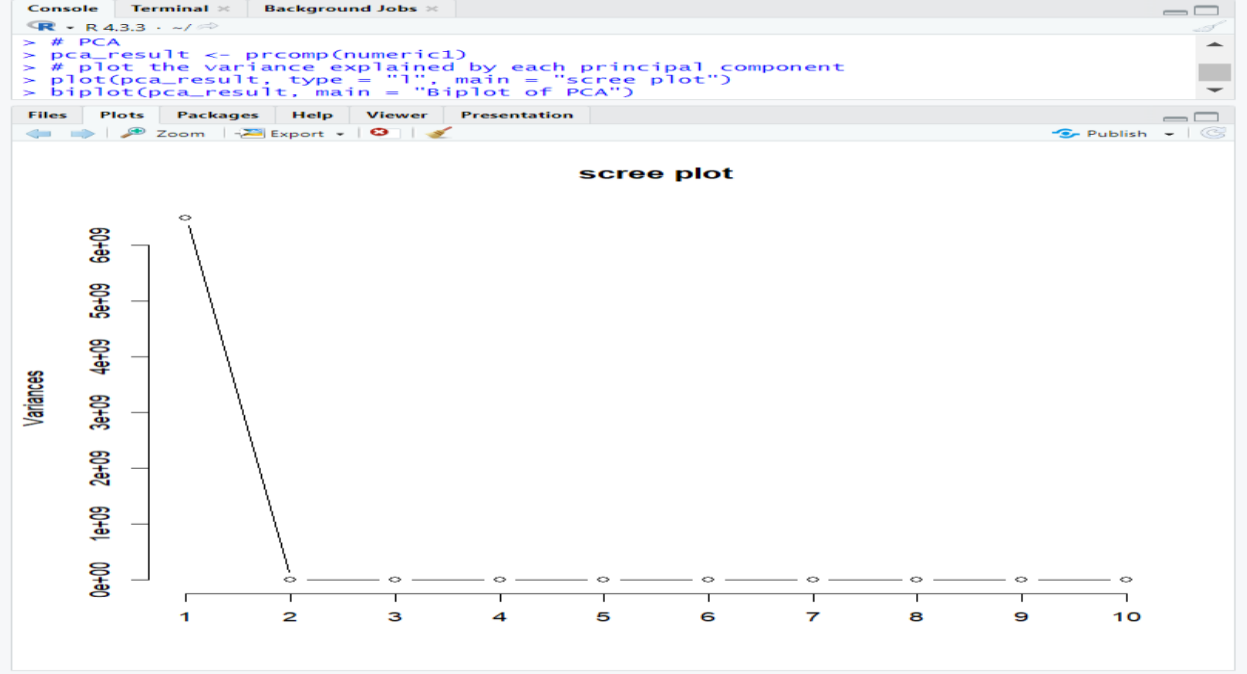
I applied ridge regression to the dataset using the same variables from the second linear model. First, I created model matrices for both the training and testing datasets. A grid of lambda values was defined, and ridge regression was fitted to the training data. Cross-validation was then performed to select the optimal lambda value, which was found to be 28.48036. Predictions were made on the test set using the best lambda, and the mean squared error (MSE) for the ridge regression model was calculated to be 6539.023.

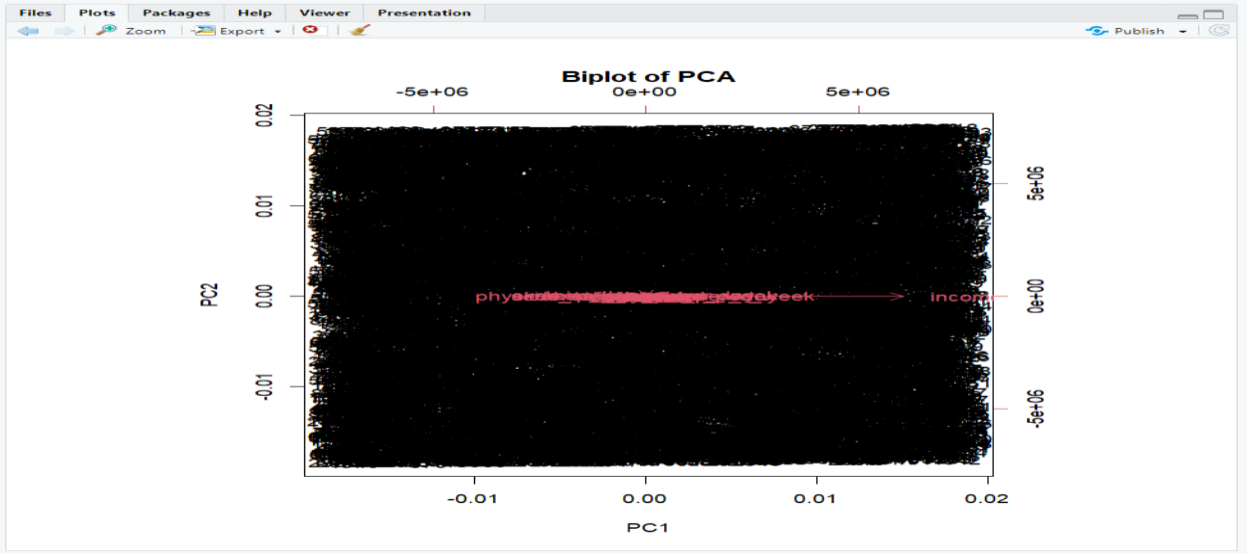
Lasso Regression

I applied lasso regression to the dataset using the same variables as in the previous models. Model matrices were created for both the training and testing datasets. A grid of lambda values was defined, and lasso regression was fitted to the training data. Cross-validation was used to select the optimal lambda value, which was determined to be 0.01. Predictions were made on the test set using this optimal lambda, and the mean squared error (MSE) for the lasso regression model was calculated to be 6543.098.

The R-squared values for the three regression models were computed to assess their predictive performance on the test dataset. The linear regression model yielded an R-squared of -0.0083, indicating that it performs worse than a simple mean prediction. Ridge regression showed a slight improvement with an R-squared of 0.0007, while lasso regression produced an R-squared of approximately 0.00003. These results suggest that none of the models explain the variability in cholesterol levels effectively, though ridge regression performs marginally better than the others.

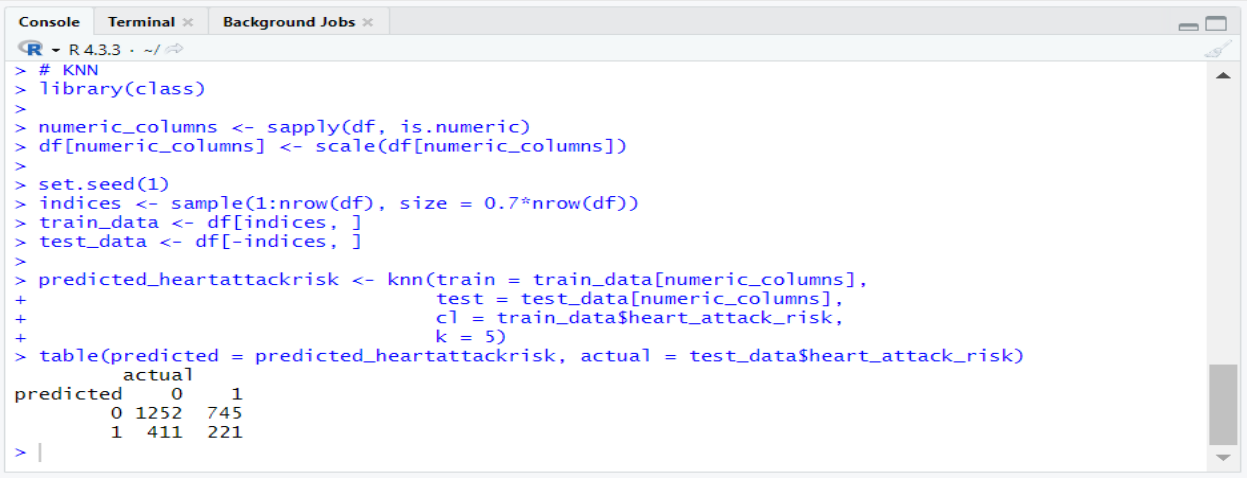
Principal Component Analysis





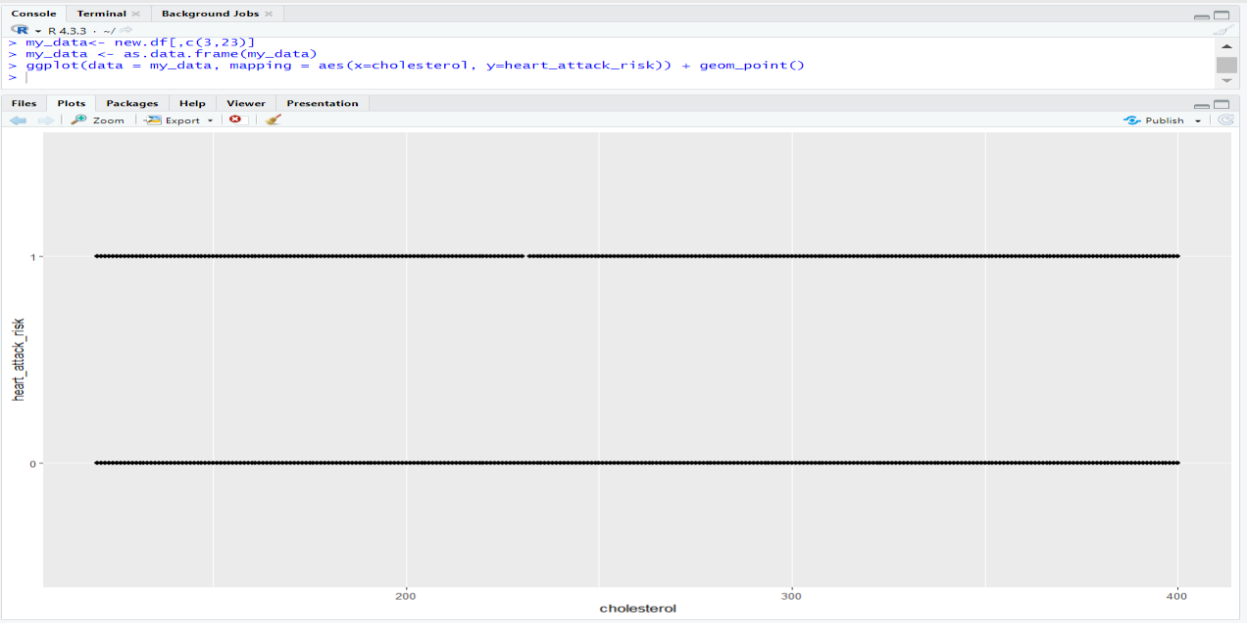
The PCA analysis indicates that the dataset has a one-dimensional structure, with the first principal component (PC1) explaining nearly all the variance. The scree plot shows a sharp decline in variance after PC1, with components 2 through 10 contributing little to no additional explanatory power. The biplot reveals that financial metrics, such as income and paycheck, strongly align with PC1, suggesting these variables are the main drivers of variation. Given the dominance of PC1, it is recommended to use this component for dimensionality reduction, retaining most of the dataset's information while minimizing complexity. The results also imply a high correlation among the original variables.

K-Nearest Neighbor

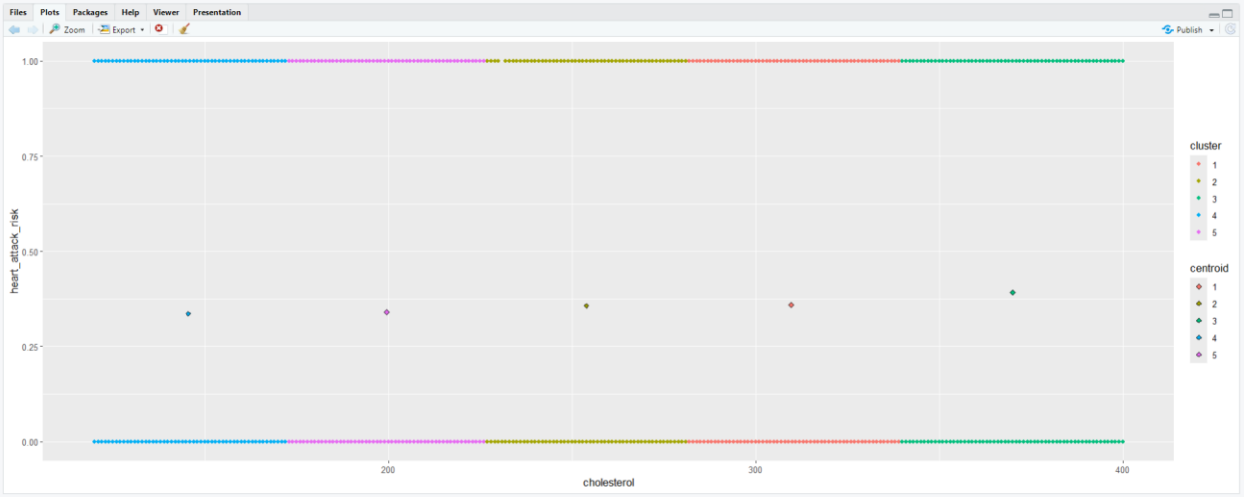


The K-Nearest Neighbors (KNN) model was trained using standardized numeric features with k=5k = 5k=5. It correctly classified 1,252 non-risk and 221 risk cases but misclassified 745 risk cases as non-risk and 411 non-risk cases as risk. The model showed a high false negative rate, indicating limited effectiveness in identifying individuals at risk.

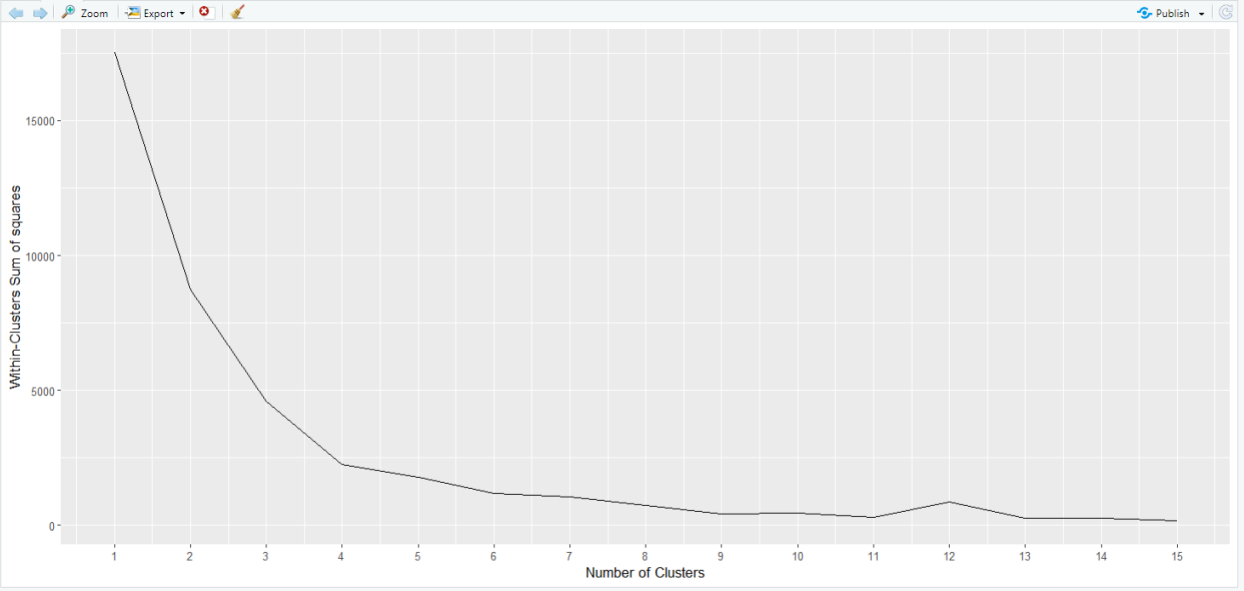
K-Means Clustering



```
> set.seed(1)
> fit <- kmeans(x=my_data, centers = 5)
> my_data$cluster <- factor(fit$cluster)
> ggplot() + geom_point(data = my_data, mapping = aes(x=cholesterol, y =heart_attack_risk, color=cluster)) + geom_point(data = data.frame(fit$centers, centroid=as.factor(1:nrow(fit$centers))), mapping = aes(x=cholesterol, y=heart_attack_risk, fill=centroid), shape=23, color="black")
> my_data1<- as.data.frame(scale(my_data[,1:2]))
> wss <- (nrow(my_data1)-1)* sum(apply(X=my_data1, MARGIN = 2, FUN = var))
>
```




```
> set.seed(1)
> for (i in 2:15) wss [i] <- sum(kmeans(my_data1, centers = i)$withinss)
> wss.df<- data.frame(cluster = 1:15, wss = wss)
> ggplot(data = wss.df, mapping = aes(x=cluster,y=wss)) + geom_line() + scale_x_continuous(breaks = seq(from=1, to=15, by=1)) + labs(x="Number of Clusters", y="Within-Clusters Sum of squares")
> km.out=kmeans (my_data1, 6, nstart=20)
> km.out$tot.withinss
[1] 960.2581
> plot(my_data1, col=(km.out$cluster +1), main= "K-Means Clustering Results with K=6", xlab="", ylab="", pch=20, cex=1)
>
```



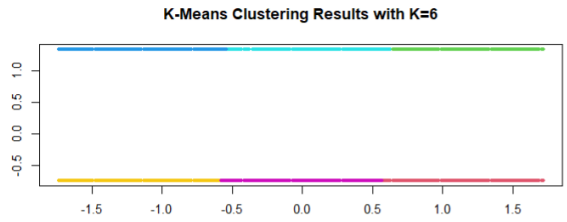
Console | Terminal | Background Jobs

R 4.3.3 · ~/

```
> set.seed(1)
> for (i in 2:15) wss [i] <- sum(kmeans(my_data1, centers = i)$withinss)
> wss.df<- data.frame(cluster = 1:15, wss = wss)
> ggplot(data = wss.df, mapping = aes(x=cluster,y=wss)) + geom_line() + scale_x_continuous(breaks = seq(from=1, to=15, by=1)) + labs(x="Number of Clusters",
y="Within-Clusters Sum of squares")
> km.out=kmeans (my_data1, 6, nstart=20)
> km.out$tot.withinss
[1] 960.2581
> plot(my_data1, col=(km.out$cluster +1), main= "K-Means Clustering Results with K=6", xlab="", ylab="", pch=20, cex=1)
>
```

Files | Plots | Packages | Help | Viewer | Presentation

Zoom | Export | Publish



SUMMARY

The Heart Attack Risk Prediction dataset was analyzed using various modeling techniques. Data preparation involved cleaning and balancing the dataset with the ovun.sample() function to address the imbalanced distribution of the target variable, heart attack risk.

Three logistic regression models were developed, with Model 2 selected as the best-performing due to its lowest test error rate. Forward and backward stepwise regression identified key predictors of heart attack risk, including cholesterol, BMI, blood pressure (normal), unhealthy diet, and alcohol consumption.

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) were applied. The LDA model achieved a test error rate of approximately 35.6%, with dietUnhealthy, alcohol consumption, and age group contributing notably to the linear discriminant. The QDA model, using predictors such as diet, cholesterol, BMI, alcohol consumption, age group, family history, and age, had prior probabilities of 64.04% for no risk and 35.96% for risk. On the test set, it correctly predicted 1,678 non-risk cases and 22 risk cases but misclassified 911 at-risk individuals as non-risk, yielding an overall error rate of 35.34%.

A Naive Bayes model trained in cholesterol, BMI, alcohol consumption, and blood pressure returned prior probabilities of 64.18% (non-risk) and 35.82% (risk). Slight group differences were observed, with cholesterol and BMI marginally higher in the risk group. Distribution patterns in alcohol consumption and blood pressure levels also varied subtly across risk categories.

Principal Component Analysis (PCA) showed that the dataset had a strong one-dimensional structure, with the first principal component (PC1) capturing nearly all the variance. A scree plot confirmed the sharp drop in explained variance after PC1. The biplot indicated that financial features like income and paycheck strongly aligned with PC1, highlighting high inter-variable correlation and supporting dimensionality reduction.

Multiple regression, ridge regression, and lasso regression were used to predict cholesterol levels. Exercise hours per week, stress levels, and heart attack risk were significant predictors. Ridge regression achieved the lowest mean squared error (MSE) at 6539.023, closely followed by lasso regression at 6543.098. However, R-squared values showed that the models explained only a small portion of the variation in cholesterol levels.

The K-Nearest Neighbors (KNN) algorithm was applied using standardized numerical features. With k=5k = 5k=5, the model correctly classified 1,252 non-risk and 221 risk cases but misclassified 745 risk cases as non-risk and 411 non-risk cases as risk, reflecting a high false negative rate.

Finally, K-Means clustering was performed to explore patterns in the standardized dataset. Clusters were visualized based on heart attack risk and cholesterol levels,

revealing natural groupings that provided insights into how individuals with similar risk profiles tend to cluster, particularly along features like BMI and diet.