Group 7

FRAMINGHAM HEART STUDY DATASET: INSIGHTS AND ANALYSIS

Overview 01

The dataset is designed to predict the 10-year risk of developing coronary heart disease (CHD). For this study, our group focused on analyzing the strength of correlation between each health attribute and the likelihood of developing CHD.

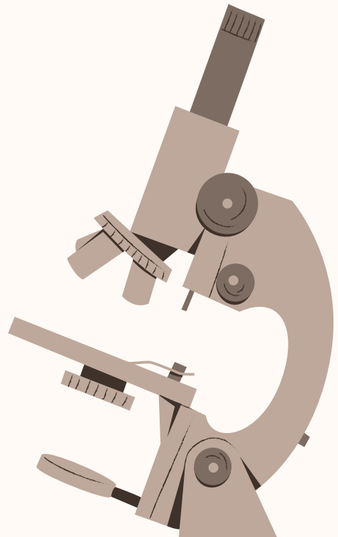
Introduction

FRAMINGHAM HEART STUDY DATASET: INSIGHTS AND ANALYSIS

The Framingham Heart Study dataset is a collection of patient records containing personal information and several health information such as heart rate, glucose level, cholesterol level, and more contributing factors.

Overview 02

- The group also categorized each variable in the dataset into one of the following four key aspects:
  - Patient Demographics: Sex, age, and education level.
  - Health Metrics: Smoking habits, blood pressure, cholesterol levels, BMI, and glucose levels.
  - Medical History: Prevalence of strokes, hypertension, and diabetes.
  - Target Variable: 10-year risk of coronary heart disease (CHD).



Data Handling

DATA IMPORT AND PREPROCESSING

```
library(dplyr)
file_path <- "framingham_heart_study.csv"
data <- read.csv(file_path, stringsAsFactors = FALSE)

data <- rename(data,sex = male)

missing_values <- sapply(data, function(x)
sum(is.na(x)))
print(missing_values)

if ("cigsPerDay" %in% colnames(data)) {
  data$cigsPerDay[is.na(data$cigsPerDay)] <- 0
}
if ("BPMeds" %in% colnames(data)) {
  data$BPMeds[is.na(data$BPMeds)] <- 0
}
if("education" %in% colnames(data)){
  data$education[is.na(data$education)] <- 0
}
columns_to_exclude_na <- c("totChol", "BMI",
"heartRate", "glucose")
data <- data[complete.cases(data[,
columns_to_exclude_na]), ]
```

Data Handling

DATA IMPORT AND PREPROCESSING

Missing Values

male	age	education	currentSmoker	cigsPerDay	BPMeds
0	0	105	0	29	53
prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP
0	0	0	50	0	0
BMI	heartRate	glucose	TenYearCHD		
19	1	388	0		

Before Cleaning

Data	
data	4240 obs. of 16 variables
Values	
file_path	"framingham_heart_study.csv"
missing_values	Named int [1:16] 0 0 105 0 29 53 0 0 0 50 ...

After Cleaning

Data	
data	3827 obs. of 16 variables
Values	
columns_to_exclude_na	chr [1:4] "totChol" "BMI" "heartRate" "glucose"
file_path	"framingham_heart_study.csv"
missing_values	Named int [1:16] 0 0 105 0 29 53 0 0 0 50 ...

Data Handling

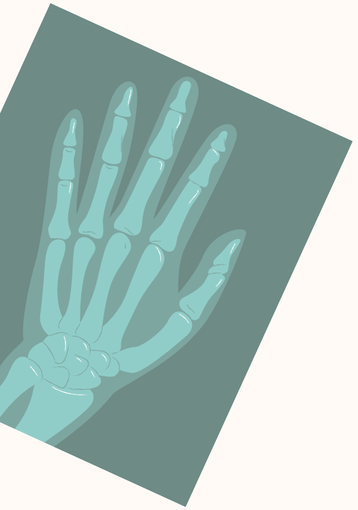
DATA EXPLORATION AND AND DESCRIPTIVE STATISTICS

```
summary(data)

str(data)
head(data)
tail(data)

numeric_cols <- sapply(data, is.numeric)
means <- sapply(data[, numeric_cols], mean, na.rm = TRUE)
print(means)

medians <- sapply(data[, numeric_cols], median, na.rm = TRUE)
print(medians)

standard_dev <- sapply(data[, numeric_cols], sd, na.rm = TRUE)
print(standard_dev)

variances <- sapply(data[, numeric_cols], var, na.rm = TRUE)
print(variances)

correlation <- cor(data[, numeric_cols], use = "complete.obs")
print(correlation)
```

```
summary_stats <- data.frame(
  Variable = names(means),
  Mean = means,
  Median = medians,
  SD = standard_dev,
  Variance = variances,
  row.names = NULL
)
summary_stats
```

Data Handling

DATA EXPLORATION AND AND DESCRIPTIVE STATISTICS



Summary

```
      male              age          education       currentSmoker
 Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000
 Median :0.0000   Median :49.00   Median :2.000   Median :0.0000
 Mean   :0.4434   Mean   :49.62   Mean   :1.978   Mean   :0.4907
 3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000
                                  NA's   :95
   cigsPerDay        BPMeds       prevalentStroke     prevalentHyp
 Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000
 1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000
 Median : 0.000   Median :0.00000   Median :0.000000   Median :0.000
 Mean   : 8.939   Mean   :0.02979   Mean   :0.005749   Mean   :0.313
 3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.000
 Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.000

    diabetes          totChol          sysBP           diaBP
 Min.   :0.00000   Min.   :113   Min.   : 83.5   Min.   : 48.00
 1st Qu.:0.00000   1st Qu.:206   1st Qu.:117.0   1st Qu.: 75.00
 Median :0.00000   Median :234   Median :128.0   Median : 82.00
 Mean   :0.02718   Mean   :237   Mean   :132.4   Mean   : 82.96
 3rd Qu.:0.00000   3rd Qu.:264   3rd Qu.:144.0   3rd Qu.: 90.00
 Max.   :1.00000   Max.   :696   Max.   :295.0   Max.   :142.50

      BMI          heartRate         glucose        TenYearCHD
 Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.0000
 1st Qu.:23.08   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.0000
 Median :25.40   Median : 75.00   Median : 78.00   Median :0.0000
 Mean   :25.81   Mean   : 75.73   Mean   : 81.91   Mean   :0.1529
 3rd Qu.:28.05   3rd Qu.: 82.00   3rd Qu.: 87.00   3rd Qu.:0.0000
 Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.0000
```

Data Handling

DATA EXPLORATION AND AND DESCRIPTIVE STATISTICS

```
> str(data)
'data.frame':   3827 obs. of  16 variables:
 $ male           : int  1 0 1 0 0 0 0 0 1 1 ...
 $ age            : int  39 46 48 61 46 43 63 45 52 43 ...
 $ education      : int  4 2 1 3 3 2 1 2 1 1 ...
 $ currentSmoker  : int  0 0 1 1 1 0 0 1 0 1 ...
 $ cigsPerDay     : num  0 0 20 30 23 0 0 20 0 30 ...
 $ BPMeds         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentHyp   : int  0 0 0 1 0 1 0 0 1 1 ...
 $ diabetes       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ totChol        : int  195 250 245 225 285 228 205 313 260 225 ...
 $ sysBP          : num  106 121 128 150 130 ...
 $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
 $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
 $ heartRate      : int  80 95 75 65 85 77 60 79 76 93 ...
 $ glucose        : int  77 76 70 103 85 99 85 78 79 88 ...
 $ TenYearCHD     : int  0 0 0 1 0 0 1 0 0 0 ...
```

```
> tail(data)
     male age education currentSmoker cigsPerDay BPMeds prevalentStroke
4234    1  50         1             1          1      0               0
4235    1  51         3             1         43      0               0
4236    0  48         2             1         20      0               0
4238    0  52         2             0          0      0               0
4239    1  40         3             0          0      0               0
4240    0  39         3             1         30      0               0
     prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose
4234            1        0     313 179.0    92 25.97        66      86
4235            0        0     207 126.5    80 19.71        65      68
4236            0        0     248 131.0    72 22.00        84      86
4238            0        0     269 133.5    83 21.47        80     107
4239            1        0     185 141.0    98 25.60        67      72
4240            0        0     196 133.0    86 20.91        85      80
     TenYearCHD
4234          1
4235          0
4236          0
4238          0
4239          0
4240          0
```

```
> head(data)
  male age education currentSmoker cigsPerDay BPMeds prevalentStroke
1    1  39         4             0          0      0               0
2    0  46         2             0          0      0               0
3    1  48         1             1         20      0               0
4    0  61         3             1         30      0               0
5    0  46         3             1         23      0               0
6    0  43         2             0          0      0               0
  prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose
1            0        0     195 106.0    70 26.97        80      77
2            0        0     250 121.0    81 28.73        95      76
3            0        0     245 127.5    80 25.34        75      70
4            1        0     225 150.0    95 28.58        65     103
5            0        0     285 130.0    84 23.10        85      85
6            1        0     228 180.0   110 30.30        77      99
  TenYearCHD
1          0
2          0
3          0
4          1
5          0
6          0
```

Data Handling

DATA EXPLORATION AND AND DESCRIPTIVE STATISTICS

data × summary_stats × correlation ×

Filter

	Variable	Mean	Median	SD	Variance
1	male	4.434283e-01	0.0	0.49685425	2.468641e-01
2	age	4.961955e+01	49.0	8.57481539	7.352746e+01
3	education	1.978296e+00	2.0	1.02256004	1.045629e+00
4	currentSmoker	4.907238e-01	0.0	0.49997927	2.499793e-01
5	cigsPerDay	8.938856e+00	0.0	11.91409059	1.419456e+02
6	BPMeds	2.978835e-02	0.0	0.17002516	2.890855e-02
7	prevalentStroke	5.748628e-03	0.0	0.07561134	5.717075e-03
8	prevalentHyp	3.130389e-01	0.0	0.46379065	2.151018e-01
9	diabetes	2.717533e-02	0.0	0.16261533	2.644374e-02
10	totChol	2.370274e+02	234.0	44.76832423	2.004203e+03
11	sysBP	1.324468e+02	128.0	22.07447079	4.872823e+02
12	diaBP	8.295898e+01	82.0	11.96564125	1.431766e+02
13	BMI	2.580720e+01	25.4	4.06685462	1.653931e+01
14	heartRate	7.573138e+01	75.0	11.93747718	1.425034e+02
15	glucose	8.190985e+01	78.0	23.75387771	5.642467e+02
16	TenYearCHD	1.528612e-01	0.0	0.35990073	1.295285e-01

Descriptive Measures

DATA HANDLING

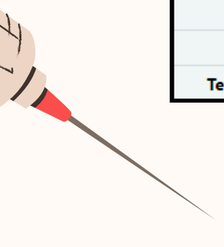
DATA EXPLORATION AND AND DESCRIPTIVE STATISTICS

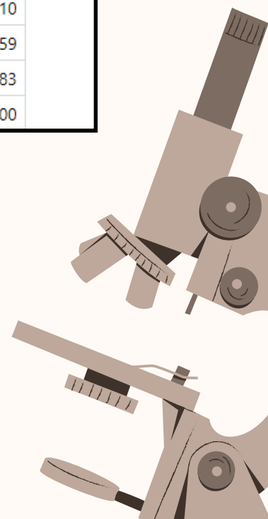
Correlation

data | summary_stats | correlation

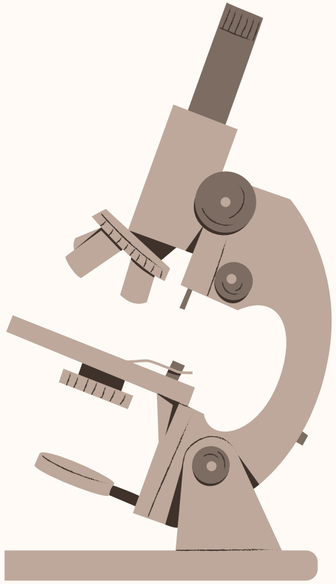
Filter

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
male	1.000000000	-0.029744558	0.02152634	0.20731360	0.32792270	-0.05096509	-0.005080413	-0.002426987	0.014540012	-0.07571803	-0.04977183	0.04898030	0.07383953	-0.115638318	0.003528572	0.08721002
age	-0.029744558	1.000000000	-0.16124944	-0.21416141	-0.19121851	0.13236698	0.055118460	0.309031947	0.108109003	0.27073998	0.39002493	0.20668748	0.13713666	-0.001969389	0.119873382	0.23507248
education	0.021526339	-0.161249435	1.00000000	0.02504990	0.01403069	-0.01325345	-0.032600362	-0.078362324	-0.041681934	-0.01511439	-0.12329953	-0.05479283	-0.13787037	-0.062279685	-0.033978369	-0.06092646
currentSmoker	0.207313601	-0.214161406	0.02504990	1.00000000	0.76558529	-0.05211857	-0.040647770	-0.110586588	-0.044945242	-0.05006187	-0.13691231	-0.11688700	-0.15705058	0.052367253	-0.054870220	0.01644963
cigsPerDay	0.327922698	-0.191218515	0.01403069	0.76558529	1.00000000	-0.04507954	-0.037349001	-0.072557389	-0.037823129	-0.03200749	-0.09724846	-0.06029273	-0.09011581	0.064094493	-0.056713136	0.05151283
BPMeds	-0.050965095	0.132366983	-0.01325345	-0.05211857	-0.04507954	1.00000000	0.110146031	0.259408866	0.048563818	0.09229602	0.26740946	0.19663583	0.10449585	0.012338207	0.053800138	0.08789998
prevalentStroke	-0.005080413	0.055118460	-0.03260036	-0.04064777	-0.03734900	0.11014603	1.000000000	0.068802547	0.008727801	0.01628099	0.06052210	0.04816936	0.03152082	-0.021870899	0.024845735	0.05487748
prevalentHyp	-0.002426987	0.309031947	-0.07836232	-0.11058659	-0.07255739	0.25940887	0.068802547	1.000000000	0.079734117	0.16785280	0.69949691	0.61735319	0.30180647	0.147471747	0.086136034	0.18249488
diabetes	0.014540012	0.108109003	-0.04168193	-0.04494524	-0.03782313	0.04856382	0.008727801	0.079734117	1.000000000	0.04387197	0.10435022	0.04890689	0.09240048	0.058559829	0.612498371	0.09444245
totChol	-0.075718034	0.270739982	-0.01511439	-0.05006187	-0.03200749	0.09229602	0.016280993	0.167852796	0.043871971	1.00000000	0.21827074	0.17205699	0.11474700	0.094926753	0.049732659	0.09632123
sysBP	-0.049771827	0.390024928	-0.12329953	-0.13691231	-0.09724846	0.26740946	0.060522101	0.699496915	0.104350217	0.21827074	1.00000000	0.78622036	0.32924172	0.186041172	0.133998052	0.22283652
diaBP	0.048980297	0.206687481	-0.05479283	-0.11688700	-0.06029273	0.19663583	0.048169364	0.617353193	0.048906892	0.17205699	0.78622036	1.00000000	0.38349834	0.182269119	0.061044453	0.14797166
BMI	0.073839526	0.137136656	-0.13787037	-0.15705058	-0.09011581	0.10449585	0.031520819	0.301806473	0.092400481	0.11474700	0.32924172	0.38349834	1.00000000	0.074854745	0.083310062	0.08177110
heartRate	-0.115638318	-0.001969389	-0.06227969	0.05236725	0.06409449	0.01233821	-0.021870899	0.147471747	0.058559829	0.09492675	0.18604117	0.18226912	0.07485474	1.000000000	0.094886542	0.01951659
glucose	0.003528572	0.119873382	-0.03397837	-0.05487022	-0.05671314	0.05380014	0.024845735	0.086136034	0.612498371	0.04973266	0.13399805	0.06104445	0.08331006	0.094886542	1.000000000	0.12229583
TenYearCHD	0.087210022	0.235072483	-0.06092646	0.01644963	0.05151283	0.08789998	0.054877482	0.182494876	0.094442451	0.09632123	0.22283652	0.14797166	0.08177110	0.019516591	0.122295827	1.00000000





Analysis



Objective

Analyze the relationships and differences between variables influencing cardiovascular health.

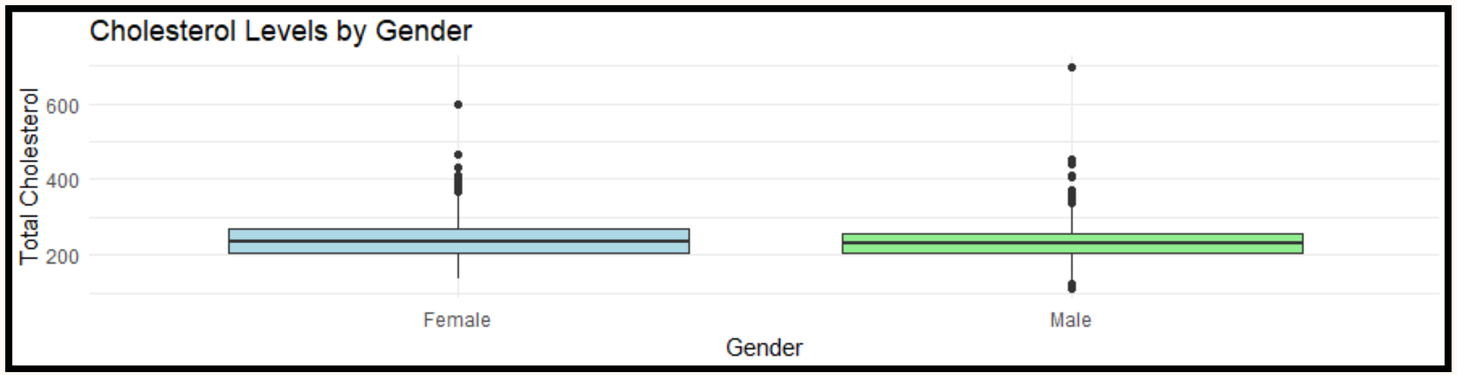
Tests Used

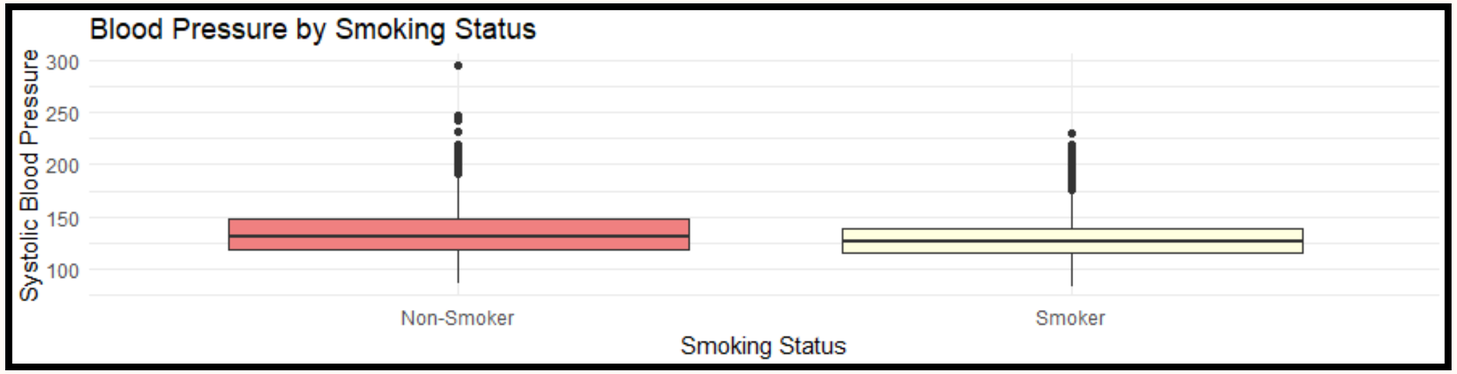
- T-Tests: Compare means of continuous variables between groups.
- Chi-Square Tests: Assess associations between categorical variables.

Applications

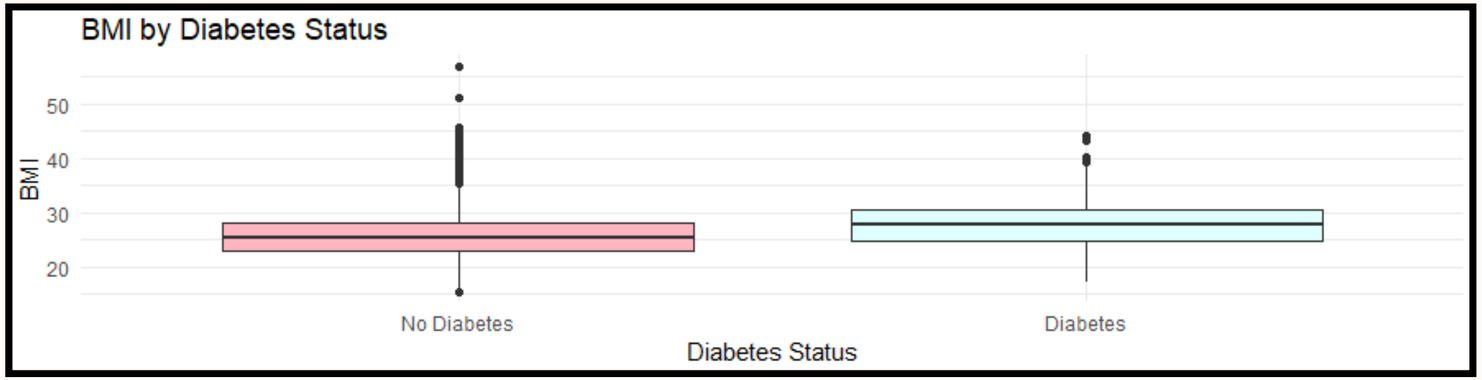
Investigate gender differences in cholesterol levels, the impact of smoking on blood pressure, the association between diabetes and heart disease, and the link between diabetes status and the occurrence of heart disease.

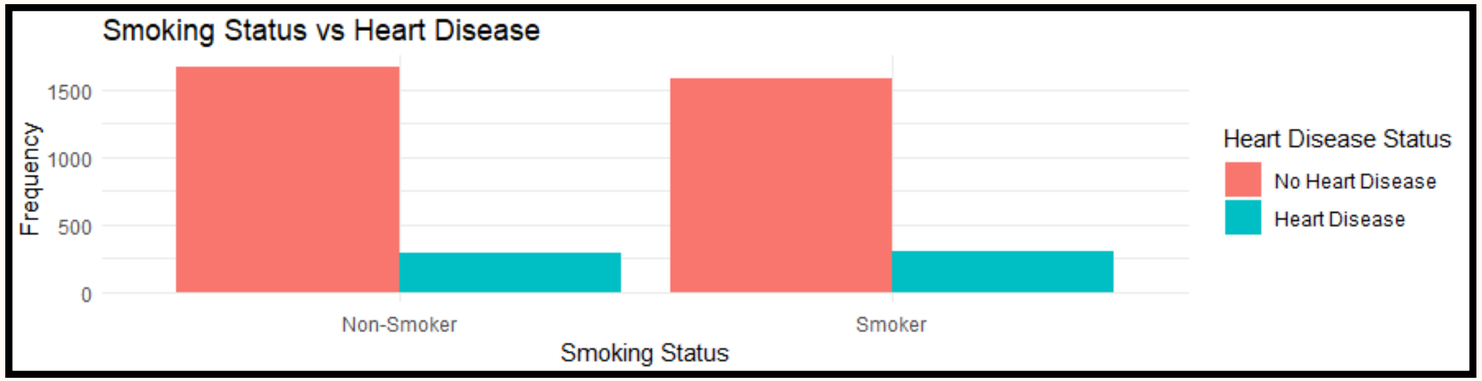
T - Tests


Cholesterol Levels by Gender
Total Cholesterol
600
400
200
Female
Male
Gender

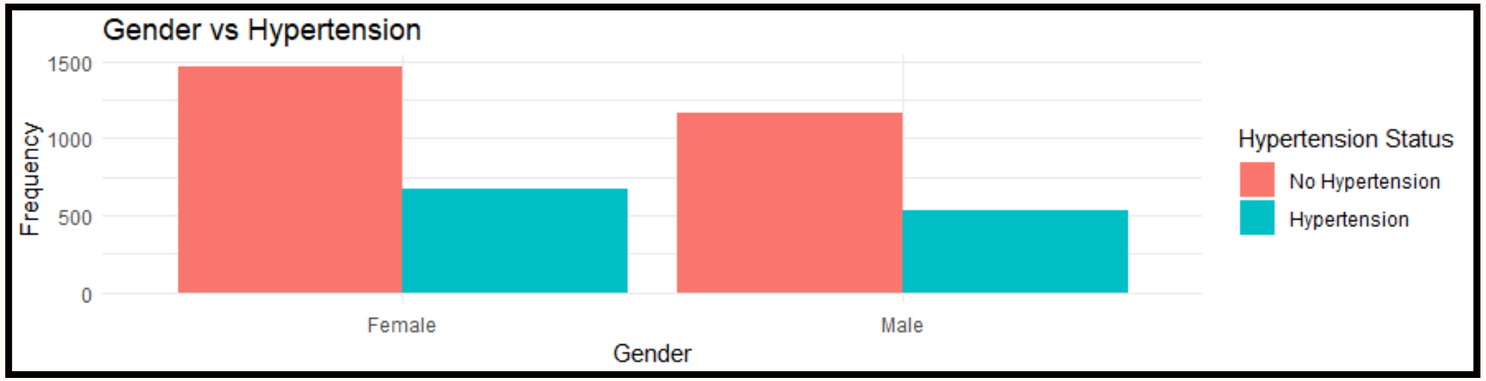

Blood Pressure by Smoking Status
Systolic Blood Pressure
300
250
200
150
100
Non-Smoker
Smoker
Smoking Status

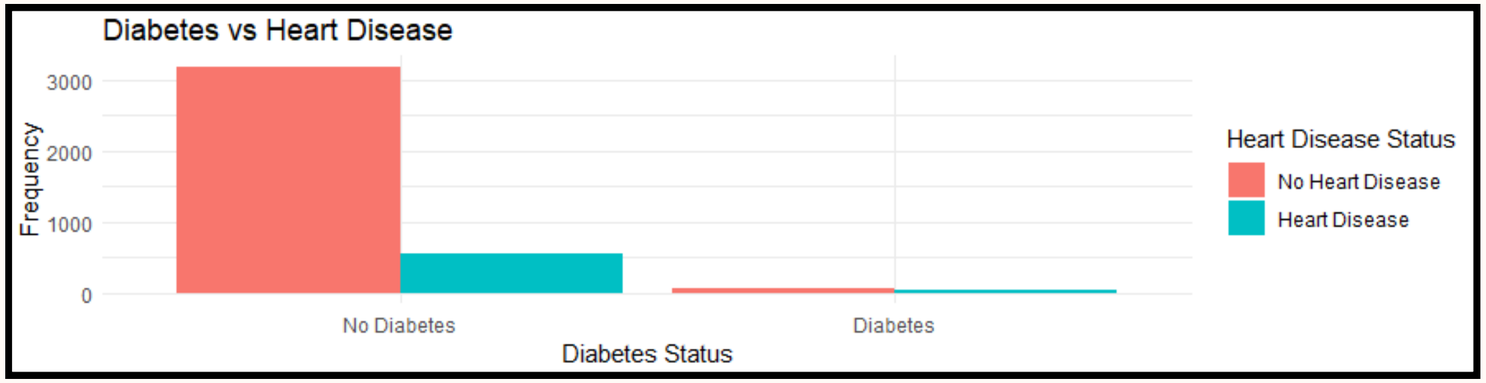
T - Tests


BMI by Diabetes Status
BMI
50
40
30
20
No Diabetes
Diabetes
Diabetes Status


Smoking Status vs Heart Disease
Frequency
1500
1000
500
0
Non-Smoker
Smoker
Smoking Status
Heart Disease Status
No Heart Disease
Heart Disease

Chi – Square Tests





Visualization Bar Graph

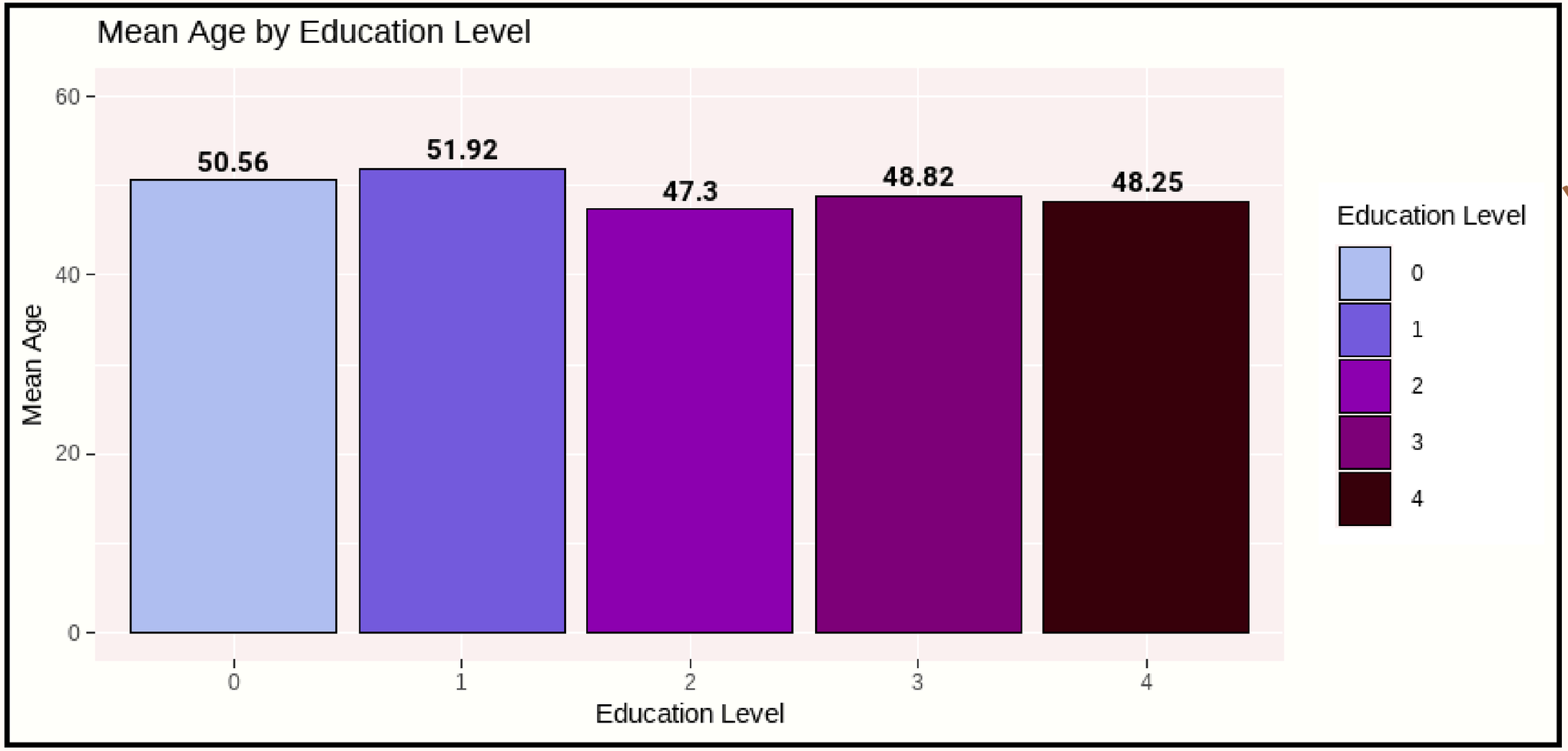
Code Overview

```
ageEduMean <- data %>%
  group_by(education) %>%
  summarize(mean_age = mean(age))  # Calculate mean age
ageEduMean$education <- as.factor(ageEduMean$education)

ggplot(ageEduMean, aes(x = education, y = mean_age, fill = education))
+
  geom_bar(stat = "identity", color = "black") +  # Use identity for
precomputed values
  labs(title = "Mean Age by Education Level", x = "Education Level", y =
"Mean Age", fill="Education Level") +
  scale_fill_manual(values = c("#B3C2F2", "#735CDD", "#9000B3",
"#7E007B", "#37000A")) +
  geom_text(aes(label = round(mean_age, 2)), vjust = -0.5, size = 4,
fontface = "bold", family = "roboto") +
  ylim(0,60) +
  theme(
    plot.background = element_rect(fill = "#FFFFFA", color = "#FFFFFA"),
    panel.background = element_rect(fill = "#FAF2F0", color =
"#FAF2F0")
  )
```

Visualization

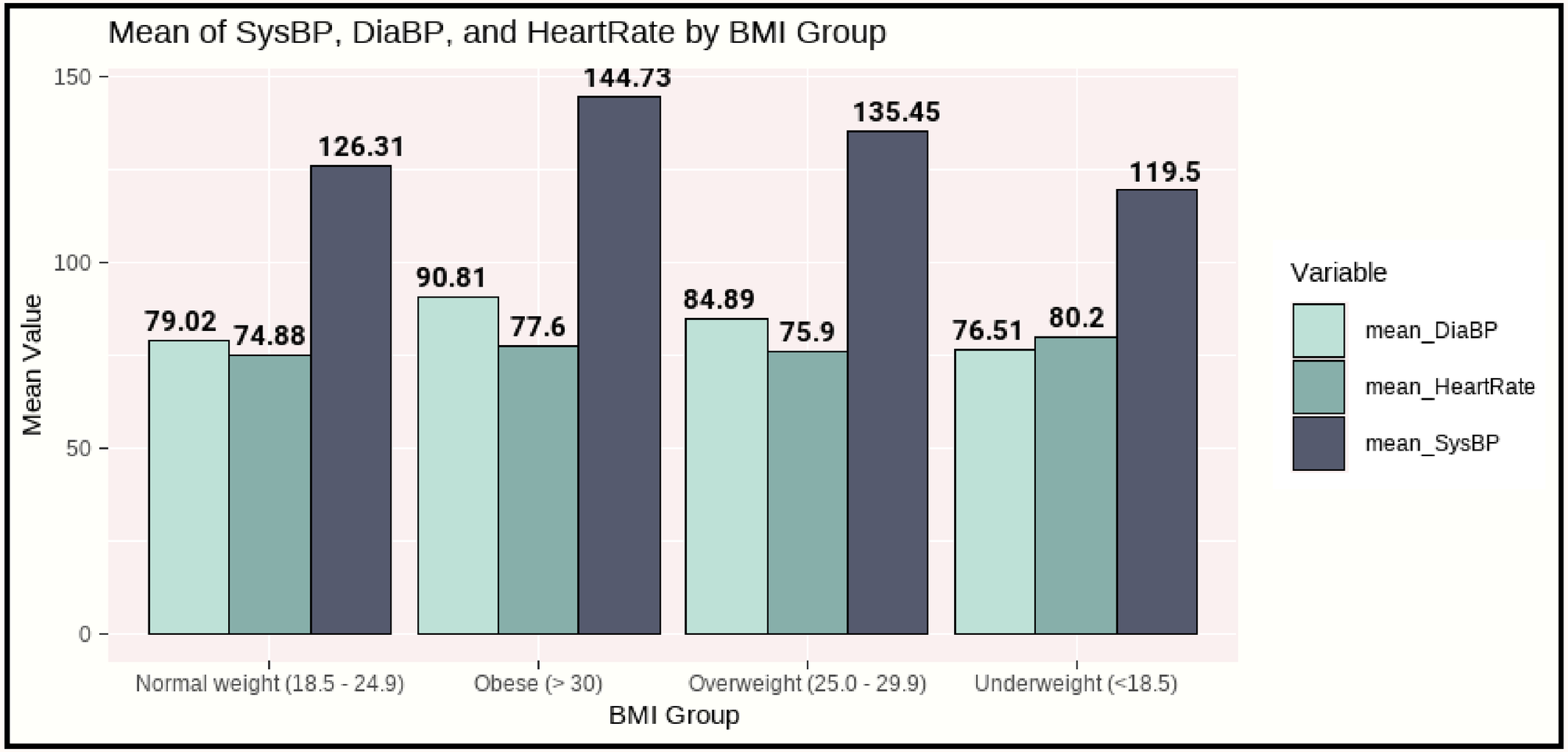
Bar Graph

Mean Age By Education level



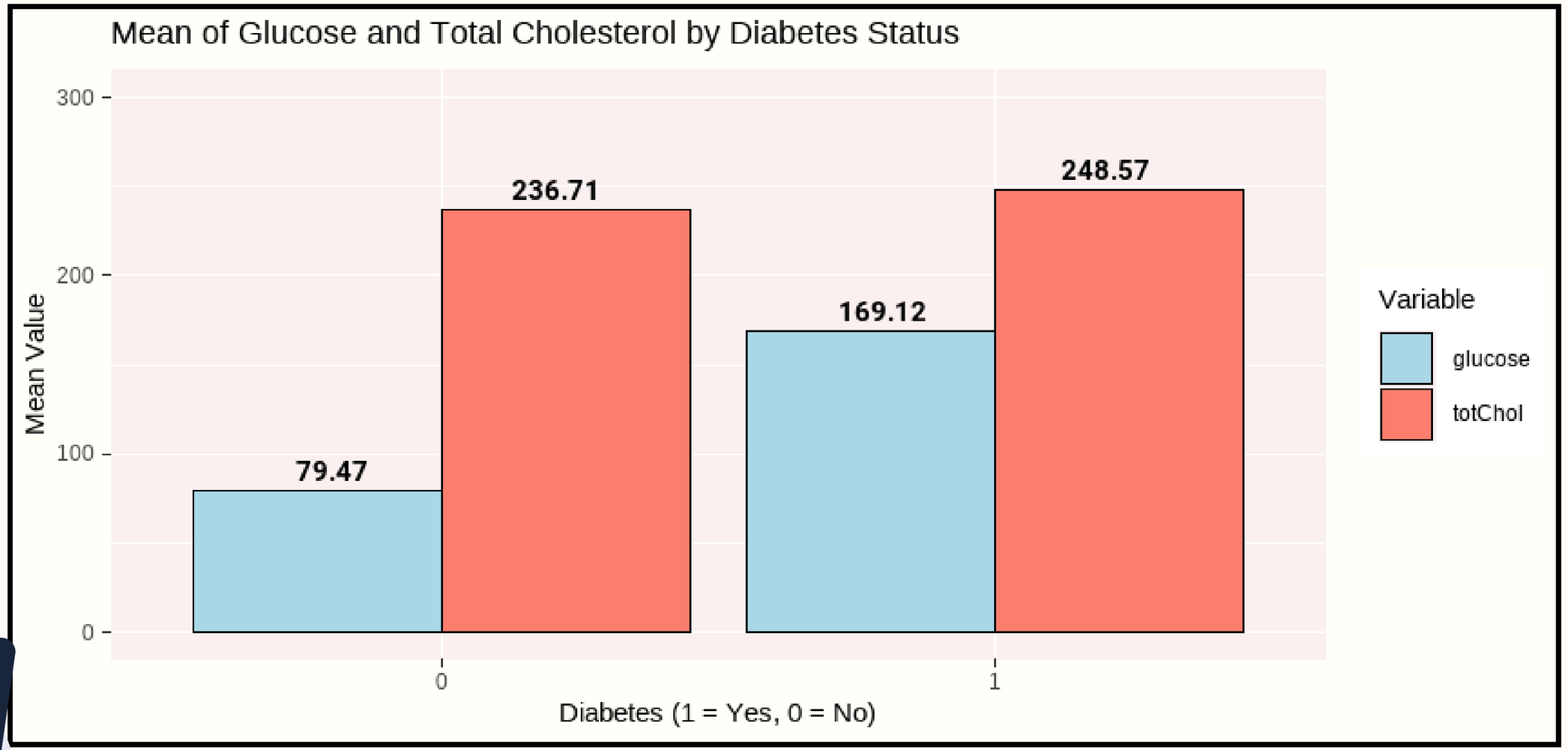
VISUALIZATION

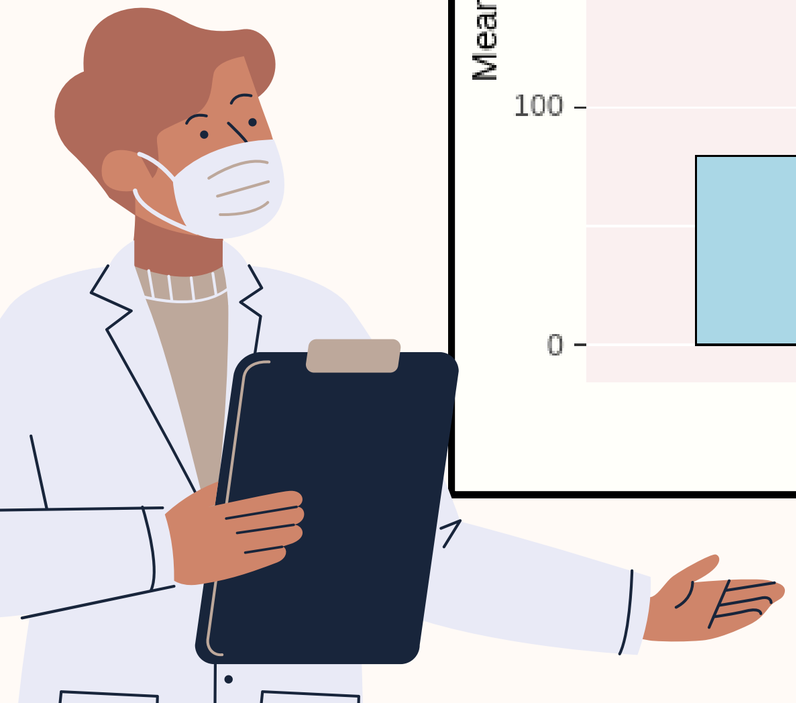
Bar Graph Mean of SysBp, DiaBP, and HeartRate by BMI Group



Visualization

Bar Graph Mean of Glucose and Total Cholesterol by Diabetes Status

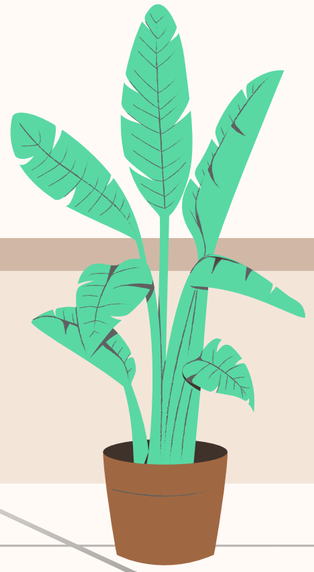




Visualization Histogram

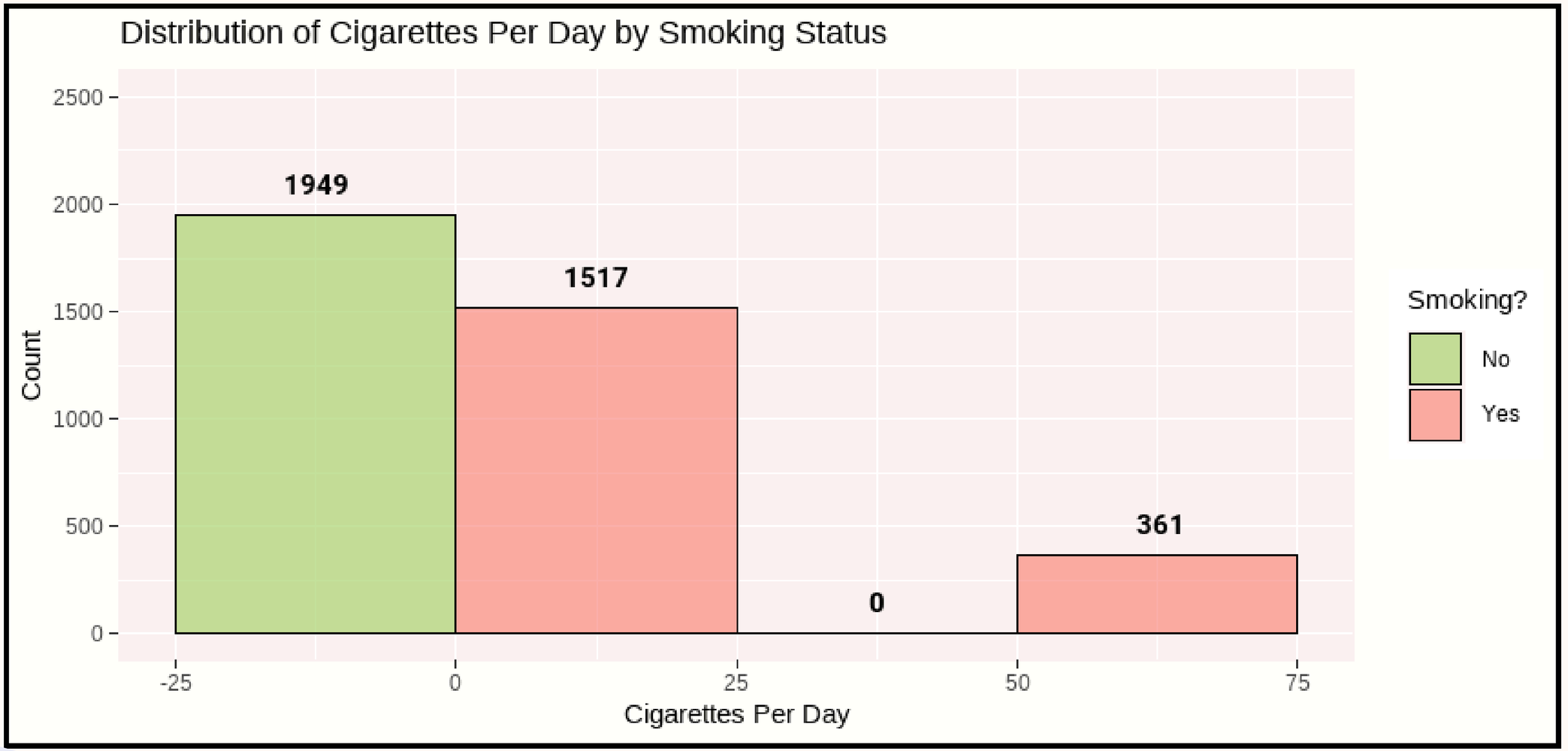
Code Overview

```
ggplot(data, aes(x = cigsPerDay, fill = factor(currentSmoker))) +
 geom_histogram(binwidth = 50, position = "dodge", color = "black",
alpha = 0.6) +
 labs(title = "Distribution of Cigarettes Per Day by Smoking Status",
     x = "Cigarettes Per Day", y = "Count", fill="Smoking?") +
 scale_fill_manual(values = c("darkolivegreen3", "salmon"),
                   labels = c("No", "Yes")) +
 stat_bin(binwidth = 50, geom = 'text', aes(label = ..count..),
          position = position_dodge(width = 50), vjust=-1.1, color = "black",
          size = 4, fontface = "bold", family = "roboto") +
 ylim(0,2500) +
 theme(
  plot.background = element_rect(fill = "#FFFFFA", color = "#FFFFFA"),
  panel.background = element_rect(fill = "#FAF2F0", color =
"#FAF2F0")
 )
```



Visualization

Histogram: Distribution of Cigarettes Per Day By Smoking Status



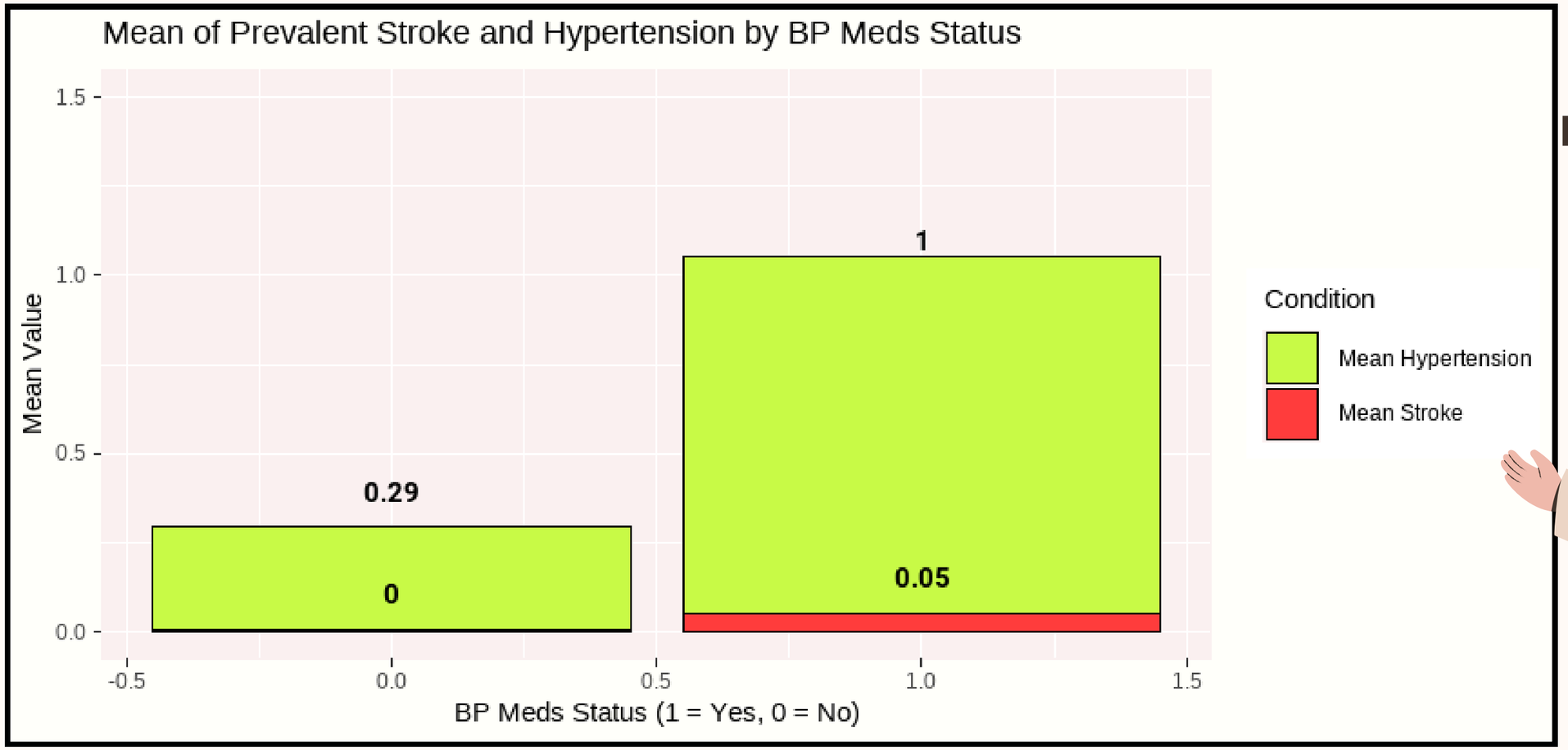
Visualization Stacked Bar Graph

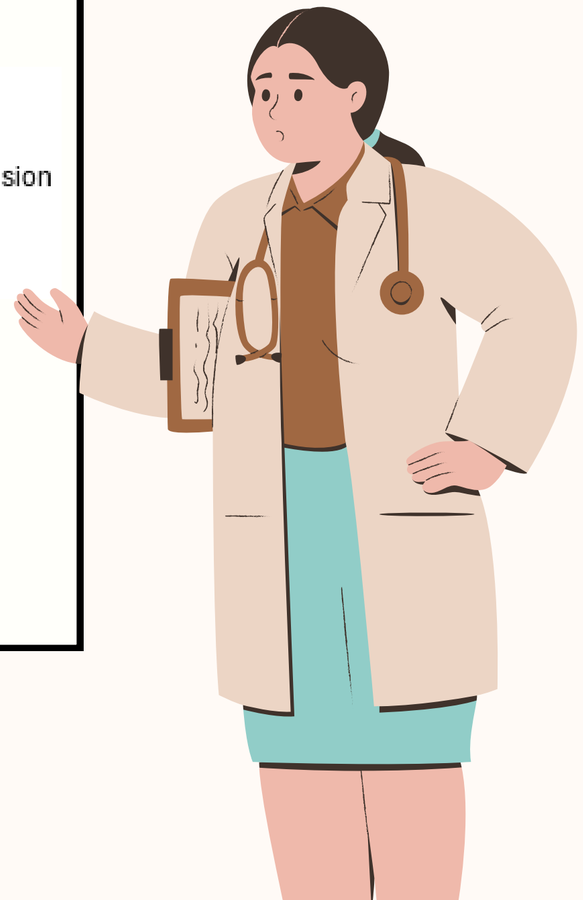
Code Overview

```
mean_data <- data %>%
 group_by(BPMeds) %>%
 summarize(
  mean_stroke = mean(prevalentStroke),
  mean_hyp = mean(prevalentHyp)
 )

mean_data_long <- mean_data %>%
 pivot_longer(cols = c(mean_hyp,mean_stroke), names_to = "Condition", values_to =
"Mean")

ggplot(mean_data_long, aes(x = BPMeds, y = Mean, fill = Condition)) +
 geom_bar(stat = "identity", position = "stack", color="black") +  # Create stacked bars
 labs(title = "Mean of Prevalent Stroke and Hypertension by BP Meds Status",
    x = "BP Meds Status (1 = Yes, 0 = No)", y = "Mean Value") +
 scale_fill_manual(values = c("mean_stroke" = "brown1", "mean_hyp" = "#CAFE48"),
            labels = c("Mean Hypertension","Mean Stroke")) +
 ylim(0,1.5) +
 geom_text(aes(label = round(Mean, 2)), position = position_nudge(y = 0.1), size =
4,fontface = "bold",family="roboto") +
theme(
  plot.background = element_rect(fill = "#FFFFFA", color = "#FFFFFA"),
  panel.background = element_rect(fill = "#FAF2F0", color = "#FAF2F0")
 )
```

Visualization

Stacked Bar Graph Mean of Prevalent Stroke and Hypertension by Bp Meds Status

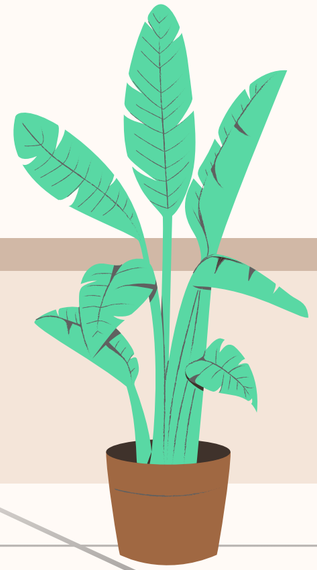




Visualization Density Plot Graph

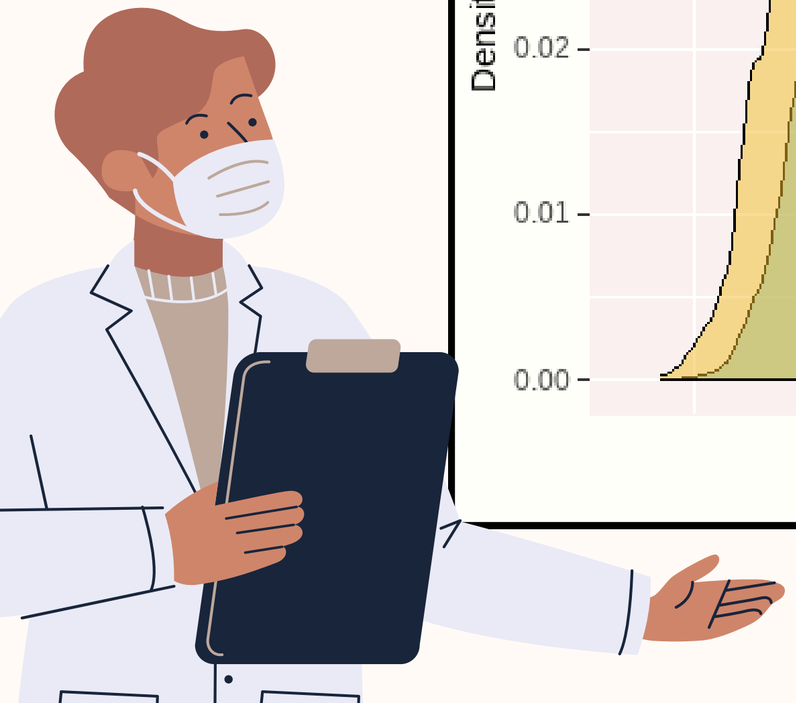
Code Overview

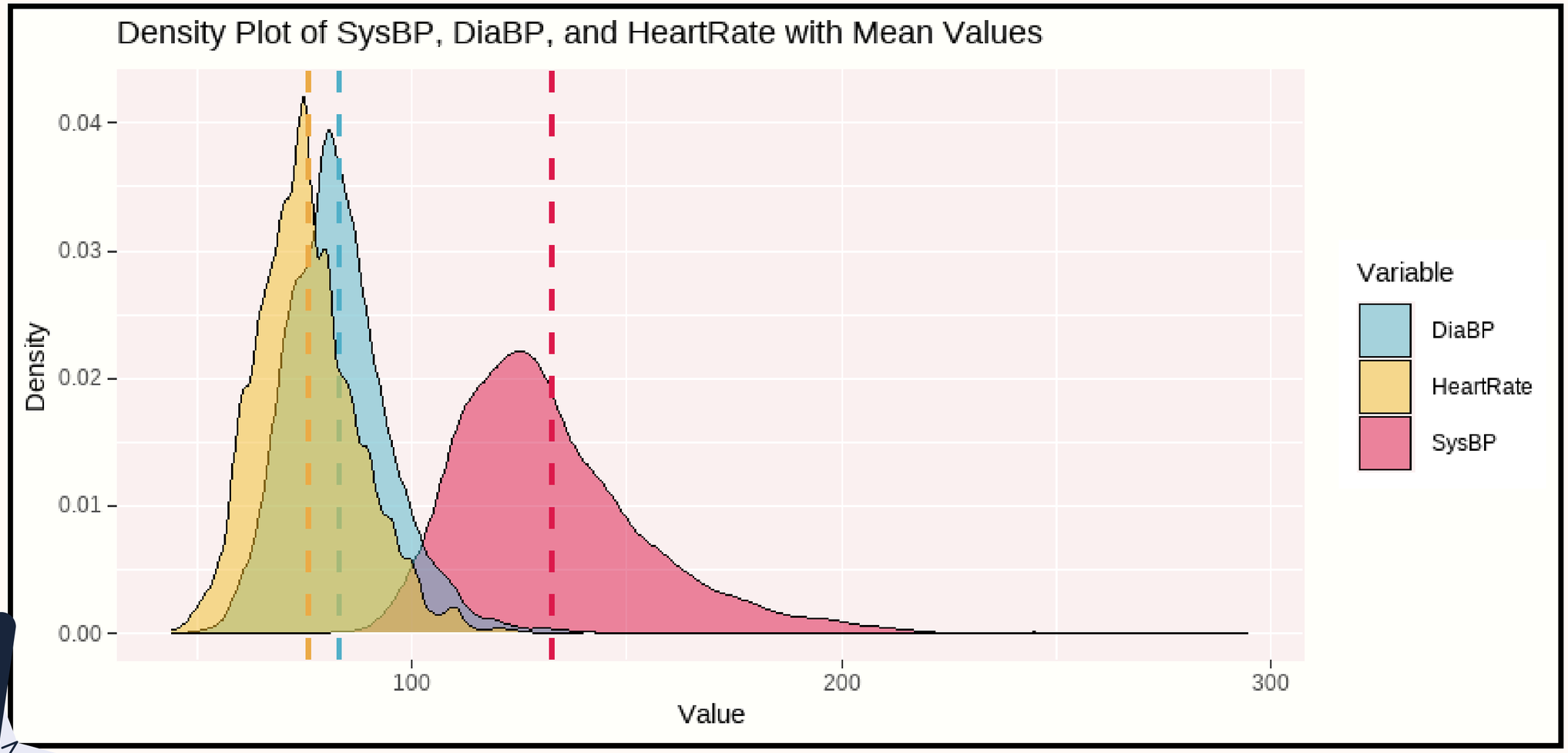
```
mean_values <- data %>%
 summarise(
  mean_SysBP = mean(sysBP, na.rm = TRUE),
  mean_DiaBP = mean(diaBP, na.rm = TRUE),
  mean_HeartRate = mean(heartRate, na.rm = TRUE)
 )

mean_SysBP <- mean_values$mean_SysBP
mean_DiaBP <- mean_values$mean_DiaBP
mean_HeartRate <- mean_values$mean_HeartRate

ggplot(data) +
 geom_density(aes(x = sysBP, fill = "SysBP"), alpha = 0.5) +  # Density plot for SysBP
 geom_vline(aes(xintercept = mean_SysBP), color = "#E01A4F", linetype = "dashed", size = 1) +
 geom_density(aes(x = diaBP, fill = "DiaBP"), alpha = 0.5) +  # Density plot for DiaBP
 geom_vline(aes(xintercept = mean_DiaBP), color = "#53B3CB", linetype = "dashed", size = 1) +
 geom_density(aes(x = heartRate, fill = "HeartRate"), alpha = 0.5) +  # Density plot for HeartRate
 geom_vline(aes(xintercept = mean_HeartRate), color = "#EDAE49", linetype = "dashed", size = 1) +
 labs(title = "Density Plot of SysBP, DiaBP, and HeartRate with Mean Values",
     x = "Value", y = "Density", fill = "Variable") +
 scale_fill_manual(values = c("SysBP" = "#E01A4F", "DiaBP" = "#53B3CB", "HeartRate" = "#F9C22E")) +
 theme(
  plot.background = element_rect(fill = "#FFFFFA", color = "#FFFFFA"),
  panel.background = element_rect(fill = "#FAF2F0", color = "#FAF2F0")
 )
```



Visualization

Density Plot of SysBp, DiaBP, and HeartRate with Mean Values





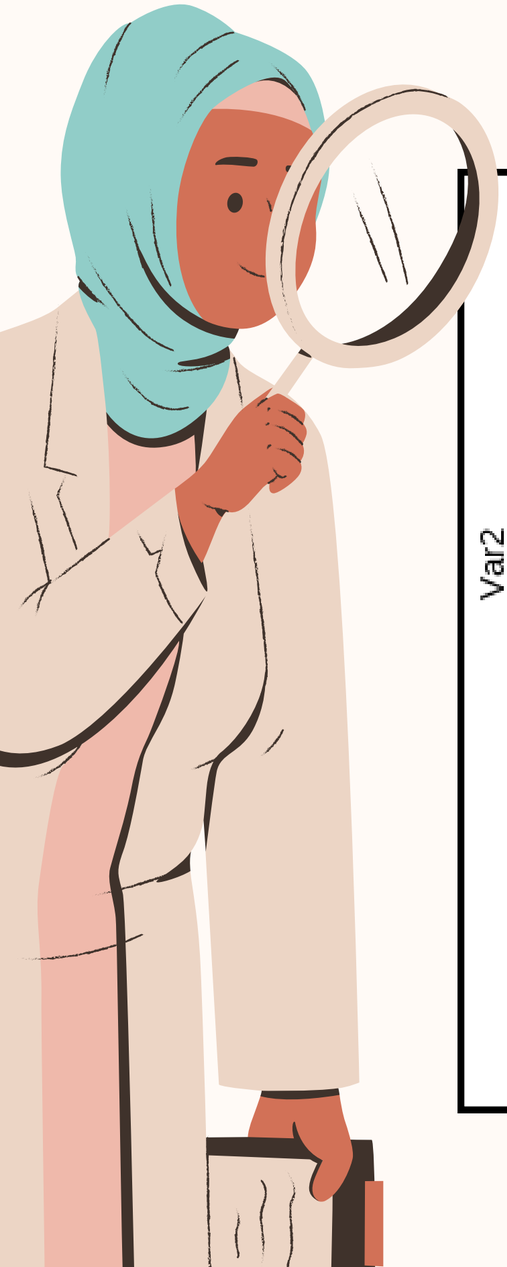
Visualization Heatmap

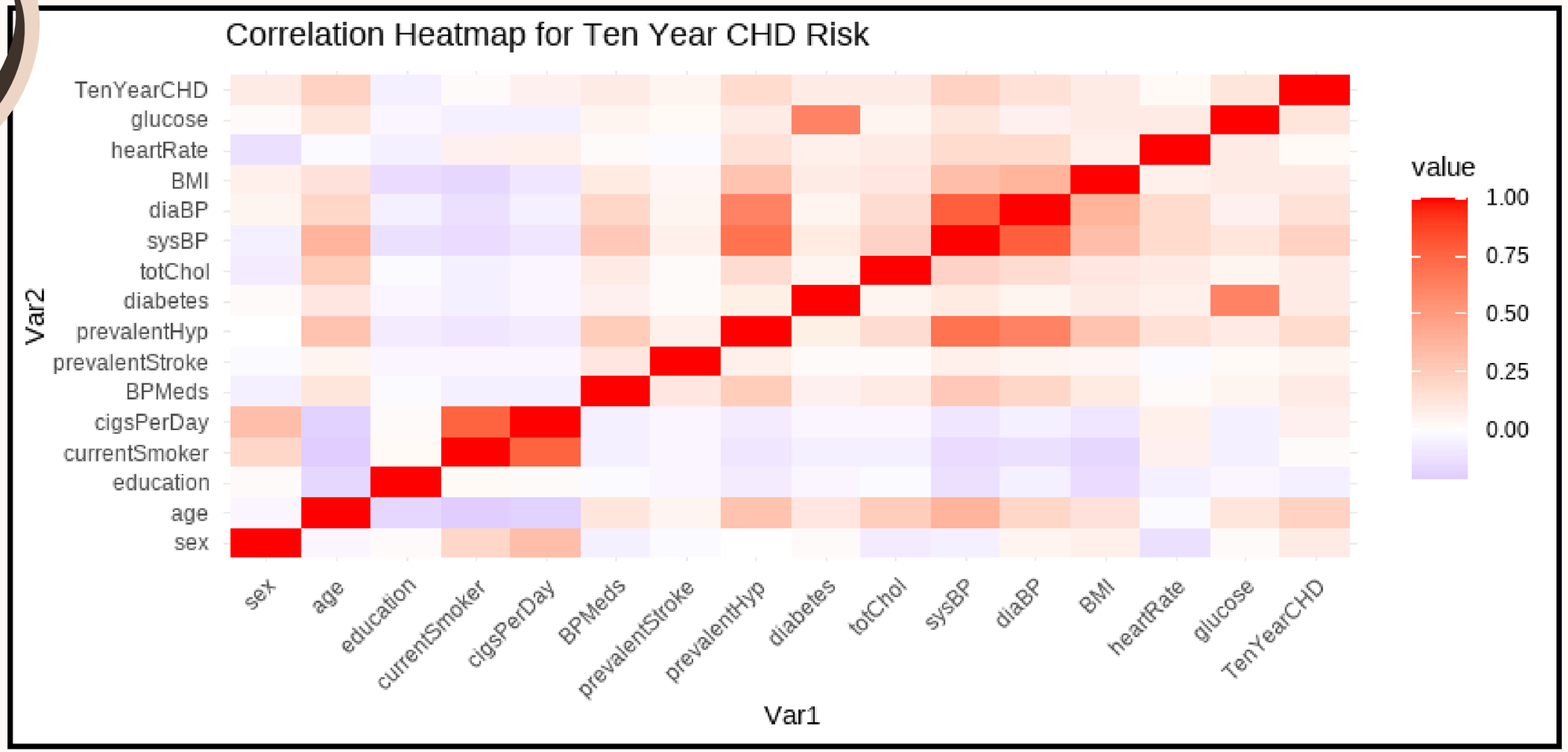
Code Overview

```
install.packages("reshape2")
library(reshape2)
corr_long <- melt(correlation)
ggplot(corr_long, aes(Var1, Var2, fill = value)) +
 geom_tile() +
 labs(title="Correlation Heatmap for Ten Year CHD Risk") +
 scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint =
0) +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
summary(data)
```

Conclusion

Correlation Heatmap for Ten year CHD Risk

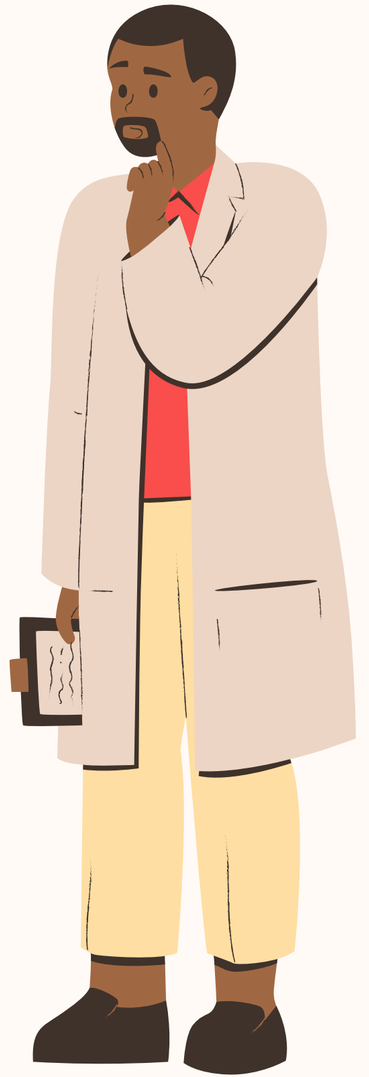




Conclusion

The analysis revealed key relationships between coronary heart disease (CHD) risk and factors like **age, hypertension, and blood pressure.** Notable correlations include **smoking history with cigarettes per day** and the link between **blood pressure metrics and hypertension or medication use**. Weak or negative correlations, such as **BMI with smoking history**, were also observed.

Key insights highlight the importance of **cholesterol levels, blood pressure, BMI, and smoking habits** in cardiovascular health. The strong link between diabetes and heart disease underscores the need for effective management of these conditions. Future research could explore **lifestyle factors and predictive modeling to enhance early interventions and targeted healthcare strategies.**



Thank you for your attention