Final_project.R

venkatasaiabhignadevarasetty

2023-04-24

```
# Load the required libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(ggplot2)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:viridis':
##
##     viridis_pal
##
```

```
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
# Read in the dataset
heart_data <- read.csv("heart_2020_cleaned.csv")

# Explore the structure and summary of the dataset
str(heart_data)
```

```
## 'data.frame':    319795 obs. of  18 variables:
##  $ HeartDisease    : chr  "No" "No" "No" "No" ...
##  $ BMI             : num  16.6 20.3 26.6 24.2 23.7 ...
##  $ Smoking         : chr  "Yes" "No" "Yes" "No" ...
##  $ AlcoholDrinking : chr  "No" "No" "No" "No" ...
##  $ Stroke          : chr  "No" "Yes" "No" "No" ...
##  $ PhysicalHealth  : num  3 0 20 0 28 6 15 5 0 0 ...
##  $ MentalHealth    : num  30 0 30 0 0 0 0 0 0 0 ...
##  $ DiffWalking     : chr  "No" "No" "No" "No" ...
##  $ Sex             : chr  "Female" "Female" "Male" "Female" ...
##  $ AgeCategory     : chr  "55-59" "80 or older" "65-69" "75-79" ...
##  $ Race            : chr  "White" "White" "White" "White" ...
##  $ Diabetic        : chr  "Yes" "No" "Yes" "No" ...
##  $ PhysicalActivity: chr  "Yes" "Yes" "Yes" "No" ...
##  $ GenHealth       : chr  "Very good" "Very good" "Fair" "Good" ...
##  $ SleepTime       : num  5 7 8 6 8 12 4 9 5 10 ...
##  $ Asthma          : chr  "Yes" "No" "Yes" "No" ...
##  $ KidneyDisease   : chr  "No" "No" "No" "No" ...
##  $ SkinCancer      : chr  "Yes" "No" "No" "Yes" ...
```

```
summary(heart_data)
```

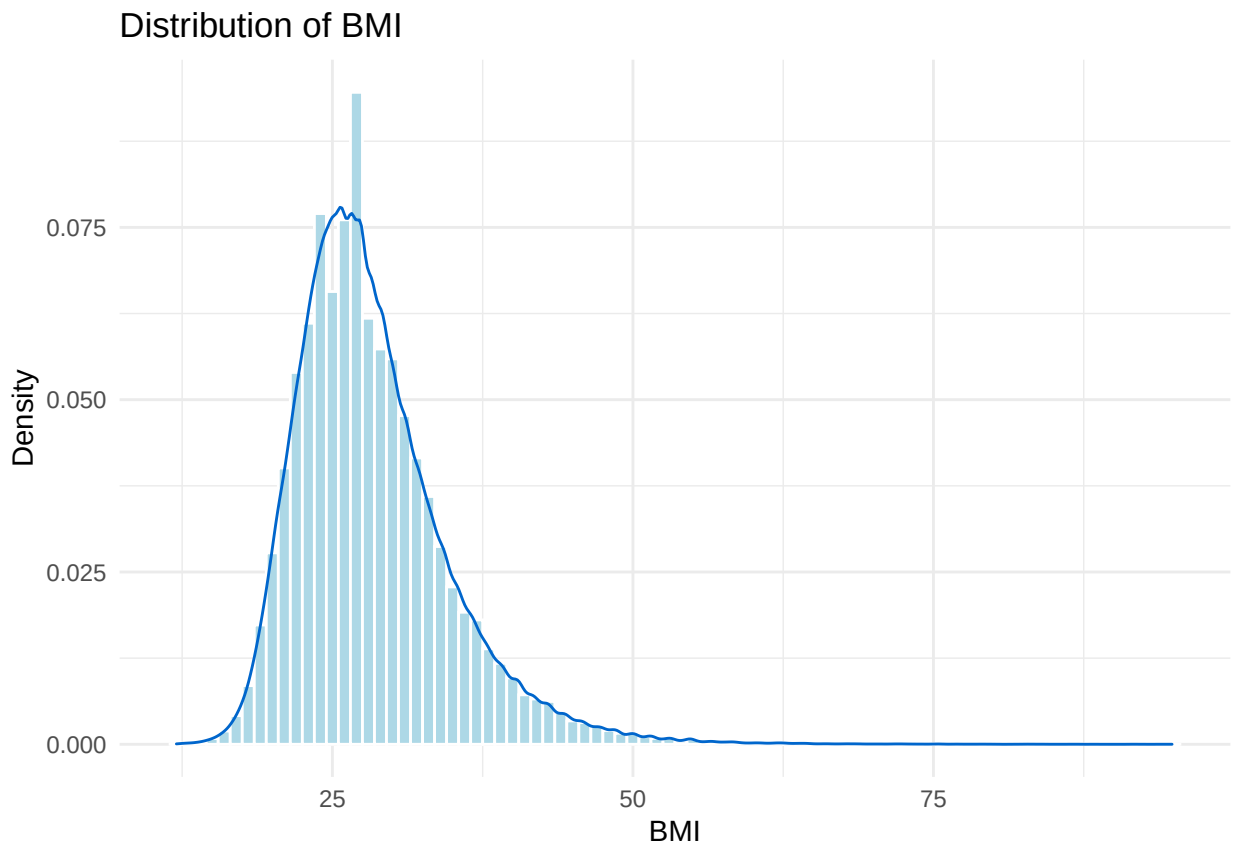
```
##  HeartDisease            BMI          Smoking          AlcoholDrinking
##  Length:319795      Min.   :12.02   Length:319795      Length:319795
##  Class :character   1st Qu.:24.03   Class :character   Class :character
##  Mode  :character   Median :27.34   Mode  :character   Mode  :character
##                     Mean   :28.33
##                     3rd Qu.:31.42
##                     Max.   :94.85
##     Stroke          PhysicalHealth    MentalHealth    DiffWalking
##  Length:319795      Min.   : 0.000   Min.   : 0.000   Length:319795
##  Class :character   1st Qu.: 0.000   1st Qu.: 0.000   Class :character
##  Mode  :character   Median : 0.000   Median : 0.000   Mode  :character
##                     Mean   : 3.372   Mean   : 3.898
##                     3rd Qu.: 2.000   3rd Qu.: 3.000
##                     Max.   :30.000   Max.   :30.000
##      Sex            AgeCategory            Race             Diabetic
##  Length:319795      Length:319795      Length:319795      Length:319795
##  Class :character   Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  PhysicalActivity    GenHealth           SleepTime         Asthma
##  Length:319795      Length:319795      Min.   : 1.000   Length:319795
##  Class :character   Class :character   1st Qu.: 6.000   Class :character
##  Mode  :character   Mode  :character   Median : 7.000   Mode  :character
##                                        Mean   : 7.097
##                                        3rd Qu.: 8.000
##                                        Max.   :24.000
##  KidneyDisease       SkinCancer
##  Length:319795      Length:319795
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

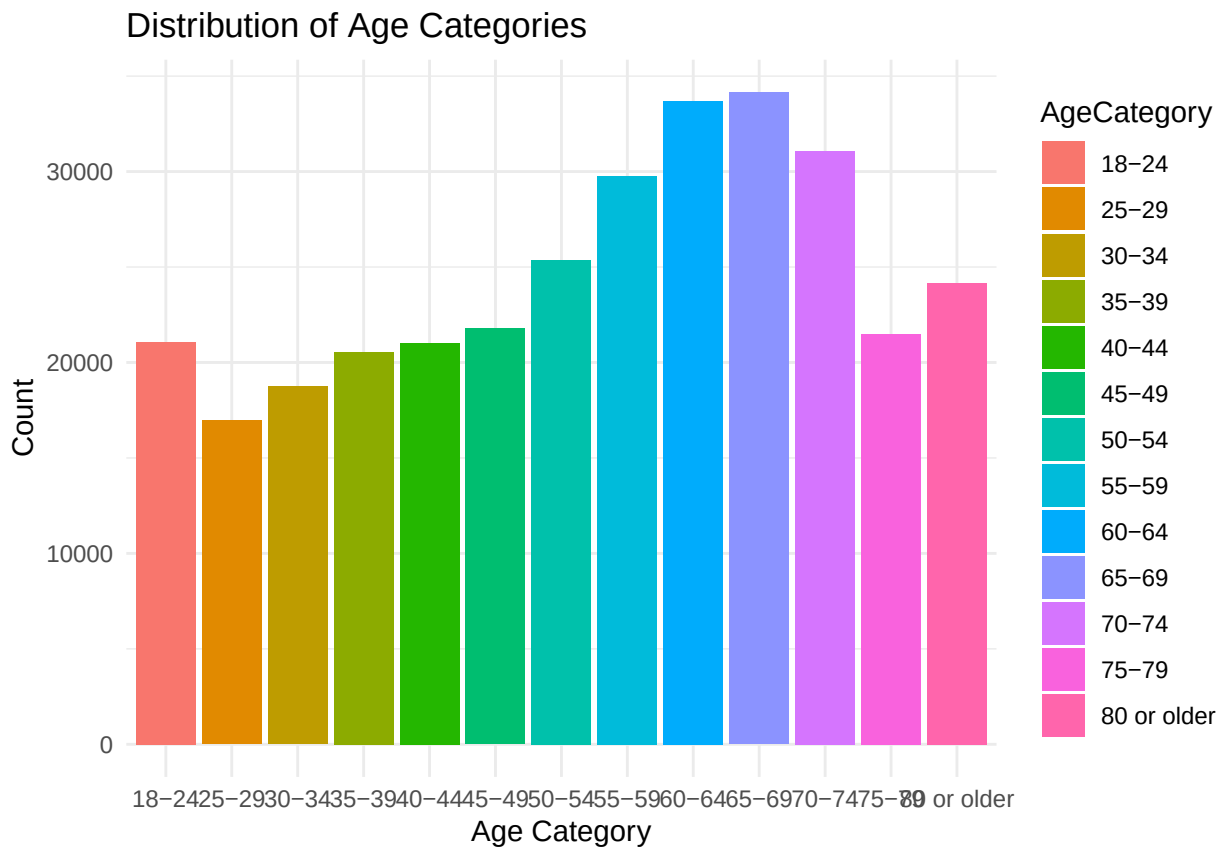
```
sum(is.na(heart_data))
```

```
## [1] 0
```

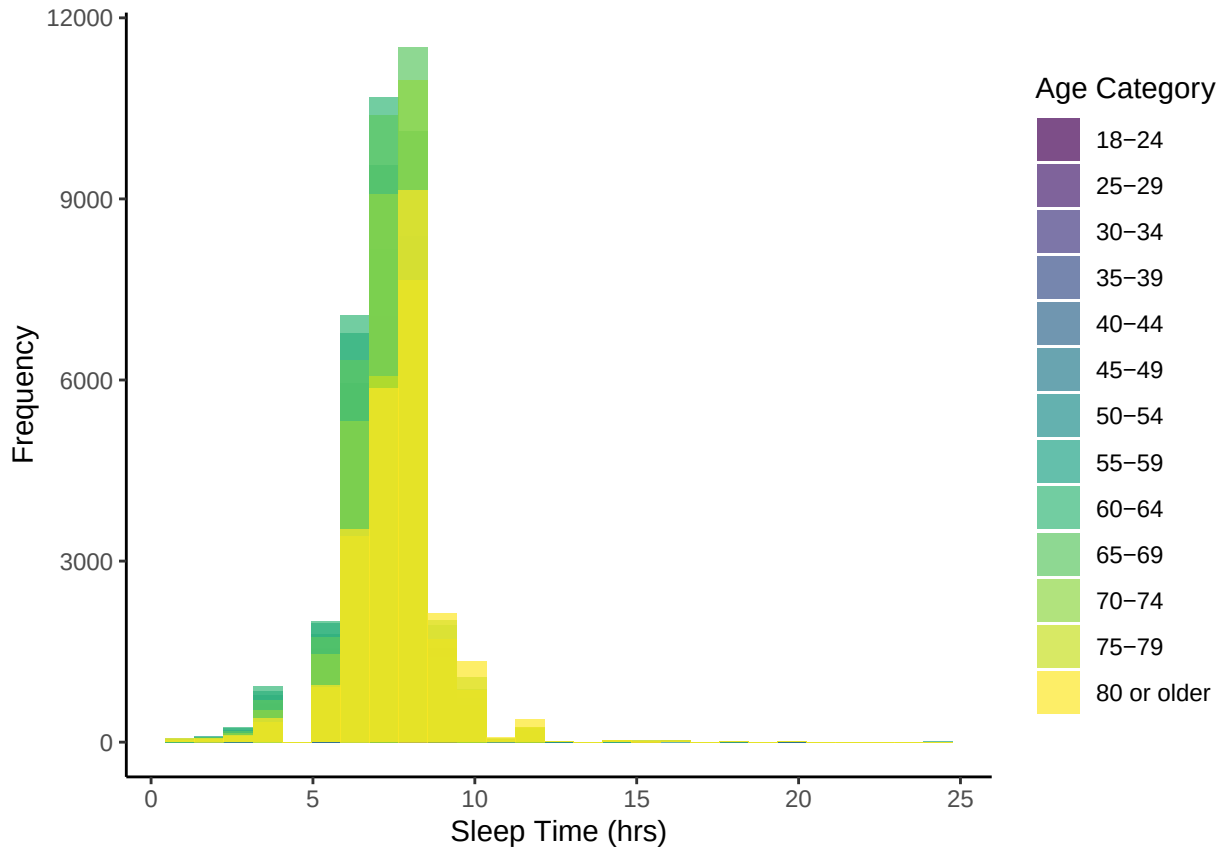
```
#The density plot shows the estimated probability density function of the BMI
#values in the dataset. The height of the curve at a given point represents the
#estimated probability that a randomly selected BMI value from the dataset will
#fall within a small range of values around that point.

#In the context of the plot you mentioned, the density on the right-hand side
#of the plot indicates that BMI values in the dataset are more concentrated in
#that range, meaning there are more individuals with BMI values in that range
#than in other ranges. Conversely, the lower density on the left-hand side of
#the plot indicates that BMI values in that range are less common in the dataset.

#Overall, the density plot provides a more continuous and detailed
#representation of the distribution of BMI values in the dataset than the
#histogram, which is based on discrete bins.


# BMI Distribution
heart_data %>%
  ggplot(aes(x = BMI, y = ..density..)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, color = "white", fill = "#ADD8E6") +
  geom_density(color = "#0066CC") +
  labs(title = "Distribution of BMI", x = "BMI", y = "Density") +
  theme_minimal()
```




```
# Distribution of Age in the data set
heart_data %>%
  count(AgeCategory) %>%
  ggplot(aes(x = AgeCategory, y = n, fill = AgeCategory)) +
  geom_bar(stat = "identity") +
  labs(title = "Distribution of Age Categories", x = "Age Category", y = "Count") +
  theme_minimal()
```

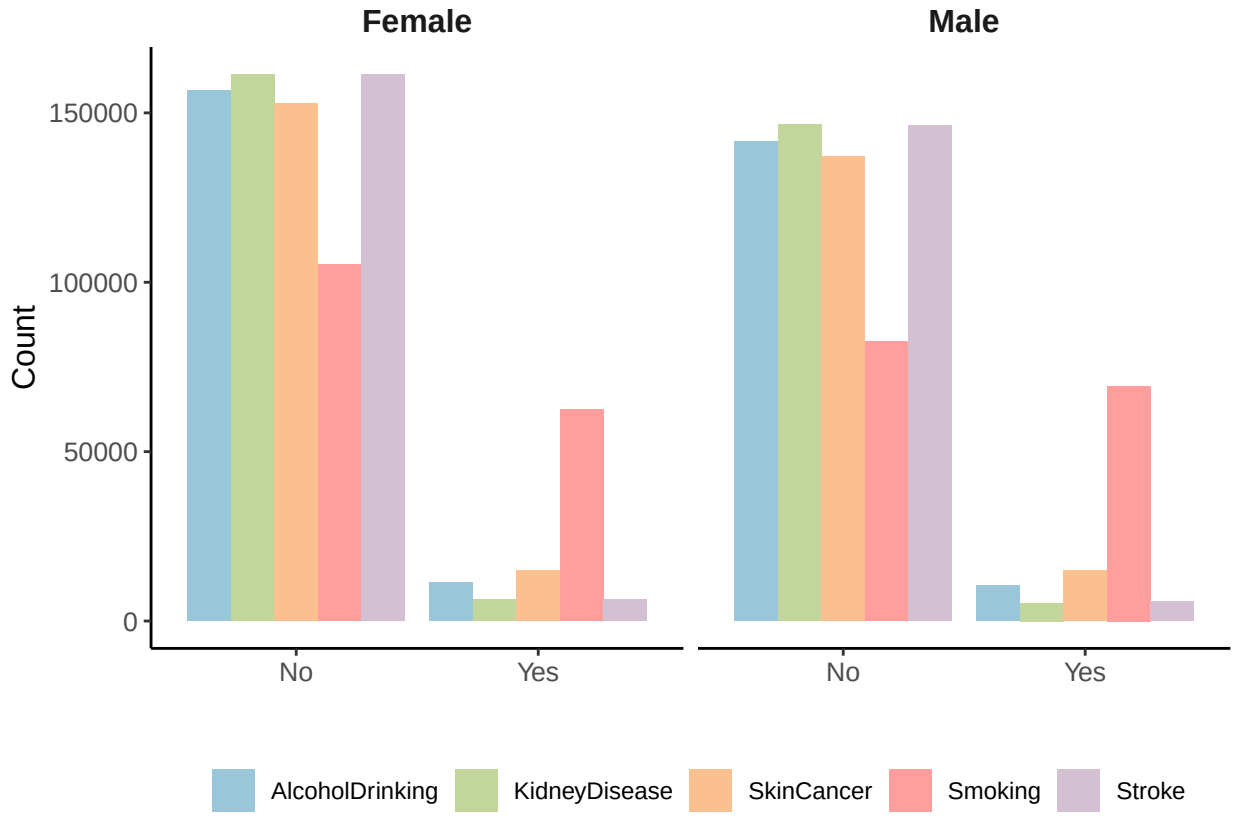



```
#Sleep Distribution Age-wise
heart_data %>%
  ggplot(aes(x = SleepTime, fill = AgeCategory)) +
  geom_histogram(binwidth = 0.9, position = "identity", alpha = .7) +
  scale_fill_viridis_d() +
  labs(x = "Sleep Time (hrs)", y = "Frequency", fill = "Age Category") +
  theme_classic()
```




```
# Ditribution of female and male across different parameters

heart_data %>%
  pivot_longer(cols = c(Smoking, AlcoholDrinking, Stroke, KidneyDisease, SkinCancer)) %>%
  group_by(name, Sex, value) %>%
  summarize(n = n()) %>%
  ggplot(aes(x = value, y = n, fill = name)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = c("#99C7D9", "#C2D69B", "#FAC090", "#FF9E9D", "#D3C0D2")) +
  facet_wrap(~ Sex, ncol = 2, scales = "free_x") +
  labs(x = "", y = "Count", fill = "") +
  theme_classic() +
  theme(strip.background = element_blank(), strip.text = element_text(size = 12, face = "bold"),
        axis.text = element_text(size = 10), axis.title = element_text(size = 12),
        legend.position = "bottom", legend.title = element_blank())
```

```
## `summarise()` has grouped output by 'name', 'Sex'. You can override using the
## `.groups` argument.
```

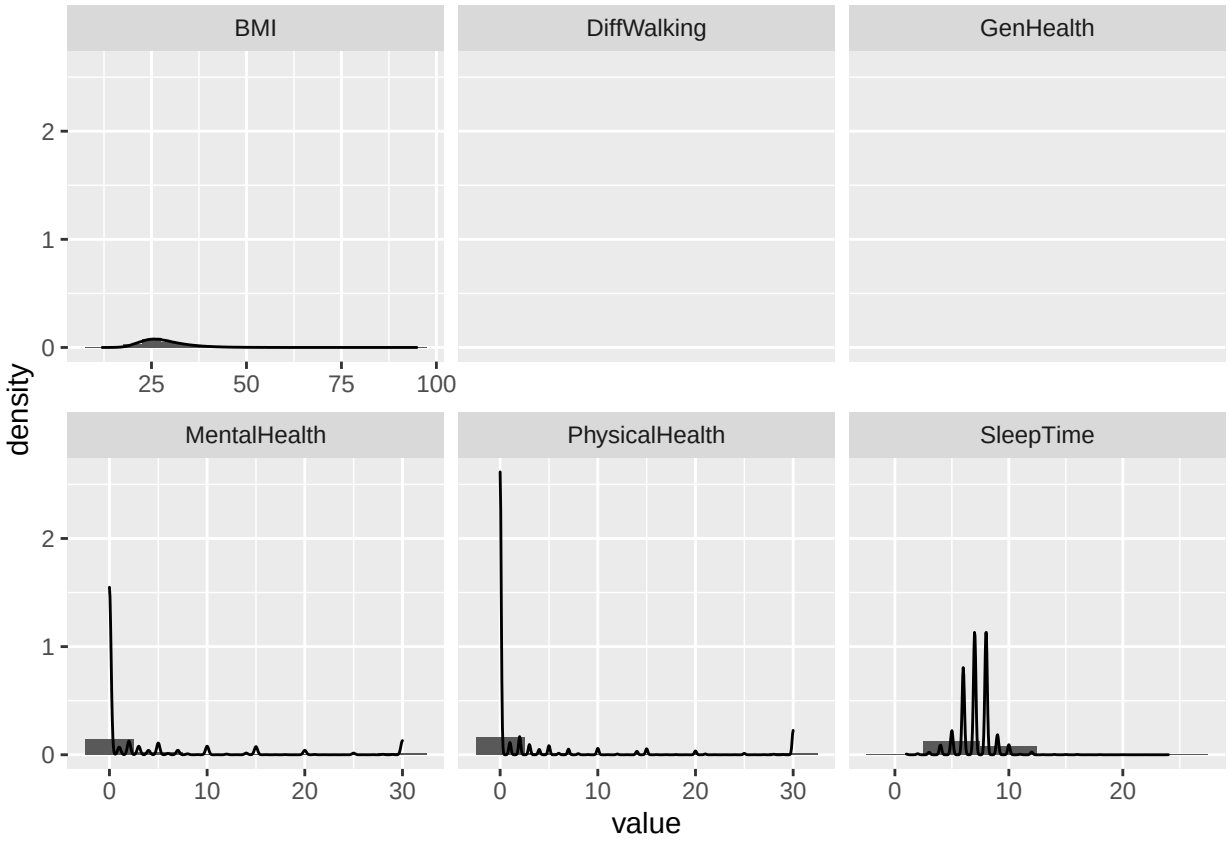



```
# Visualize the distribution of continuous variables using histograms
heart_data %>%
  select(BMI, DiffWalking, GenHealth, MentalHealth, PhysicalHealth, SleepTime) %>%
  gather(key = "variable", value = "value") %>%
  mutate(value = as.numeric(value)) %>% # Convert the value variable to numeric
  ggplot(aes(x = value, y = ..density..)) +
  facet_wrap(~variable, scales = "free_x") +
  geom_histogram(aes(y = ..density..), binwidth = 5) +
  geom_density()
```

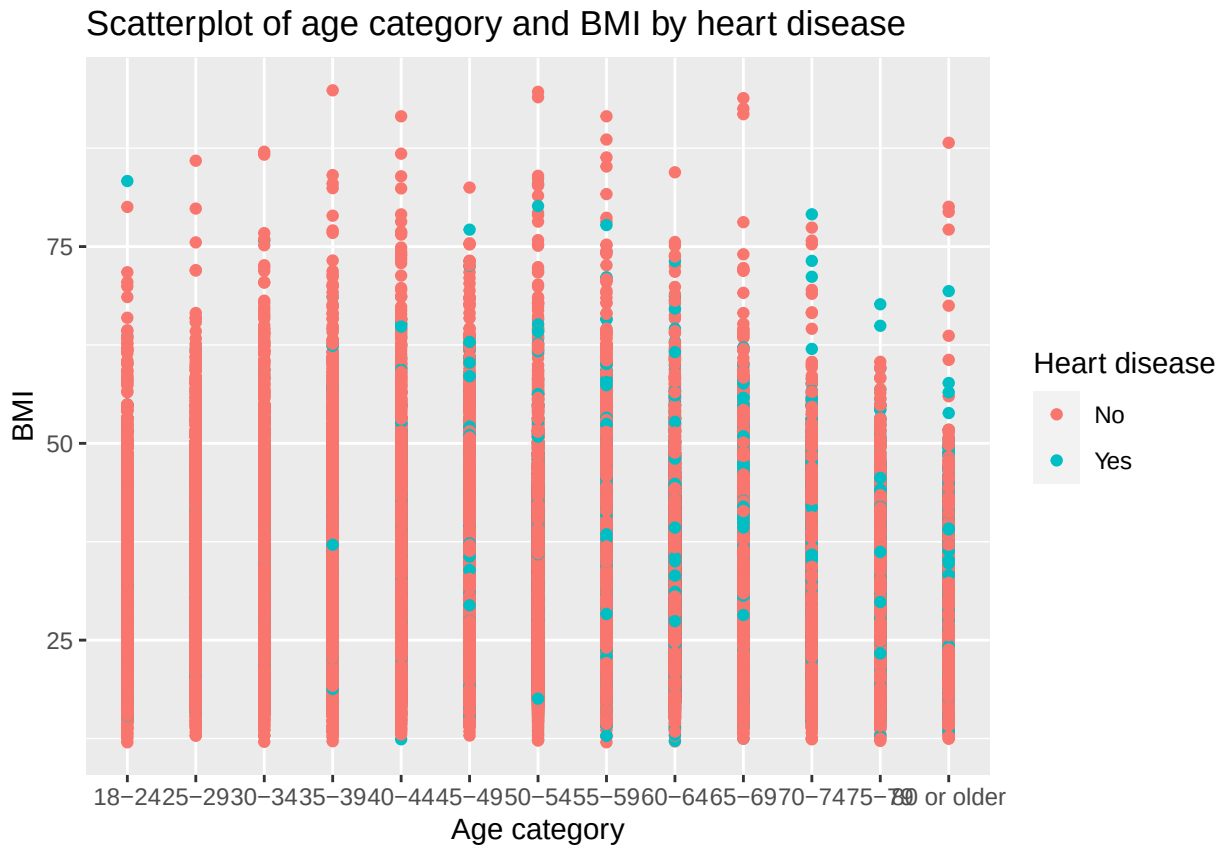
```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
## Warning: Removed 639590 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 639590 rows containing non-finite values (stat_density).
```




```
# Visualize the relationship between continuous variables using scatterplots
heart_data %>%
  select(AgeCategory, BMI, DiffWalking, GenHealth, MentalHealth, PhysicalHealth, SleepTime, HeartDiseas
  ggplot(aes(x = AgeCategory, y = BMI, color = HeartDisease)) +
  geom_point() +
  labs(title = "Scatterplot of age category and BMI by heart disease",
       x = "Age category", y = "BMI", color = "Heart disease")
```




```
# Perform t-tests to compare the means of continuous variables by the presence of heart disease

#The p-value is less than 2.2e-16, which indicates that there is strong evidence
#to reject the null hypothesis that the means of BMI in individuals with and
#without heart disease are equal.
t.test(heart_data$BMI ~ heart_data$HeartDisease)
```

```
##
##  Welch Two Sample t-test
##
## data:  heart_data$BMI by heart_data$HeartDisease
## t = -28.402, df = 32295, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -1.258156 -1.095712
## sample estimates:
##  mean in group No mean in group Yes
##          28.22466          29.40159
```

```
#we can reject the null hypothesis which states that there is no significant
#difference in the mean MentalHealth score between the groups with and without heart disease.
t.test(heart_data$MentalHealth ~ heart_data$HeartDisease)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  heart_data$MentalHealth by heart_data$HeartDisease
## t = -14.189, df = 31219, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.9252877 -0.7006833
## sample estimates:
##  mean in group No mean in group Yes
##          3.828778          4.641764
```

```
#we can reject the null hypothesis which states that there is no significant
#difference in the mean PhysicalHealth score between the groups with and without heart disease.
t.test(heart_data$PhysicalHealth ~ heart_data$HeartDisease)
```

```
##
##  Welch Two Sample t-test
##
## data:  heart_data$PhysicalHealth by heart_data$HeartDisease
## t = -68.557, df = 29536, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -4.990539 -4.713113
## sample estimates:
##  mean in group No mean in group Yes
##          2.956416          7.808242
```

```
#we can reject the null hypothesis which states that there is no significant
#difference in the mean Sleeptime score between the groups with and without heart disease.
t.test(heart_data$SleepTime ~ heart_data$HeartDisease)
```

```
##
##  Welch Two Sample t-test
##
## data:  heart_data$SleepTime by heart_data$HeartDisease
## t = -3.8607, df = 30618, p-value = 0.0001133
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.06443848 -0.02104092
## sample estimates:
##  mean in group No mean in group Yes
##          7.093416          7.136156
```

```
# Perform chi-squared tests to compare the proportions of categorical variables
#by the presence of heart disease

#very small p-value (less than 2.2e-16). This indicates strong evidence against
# the null hypothesis, and suggests that there is a significant association between
#smoking and heart disease in the sample.

#likewise for the rest of the variables as well

chisq.test(heart_data$Smoking, heart_data$HeartDisease)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  heart_data$Smoking and heart_data$HeartDisease
## X-squared = 3713, df = 1, p-value < 2.2e-16
```

```
chisq.test(heart_data$AlcoholDrinking, heart_data$HeartDisease)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  heart_data$AlcoholDrinking and heart_data$HeartDisease
## X-squared = 328.65, df = 1, p-value < 2.2e-16
```

```
chisq.test(heart_data$Stroke, heart_data$HeartDisease)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  heart_data$Stroke and heart_data$HeartDisease
## X-squared = 12386, df = 1, p-value < 2.2e-16
```

```
chisq.test(heart_data$Sex, heart_data$HeartDisease)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  heart_data$Sex and heart_data$HeartDisease
## X-squared = 1568.3, df = 1, p-value < 2.2e-16
```

```
chisq.test(heart_data$Race, heart_data$HeartDisease)
```

```
##
##  Pearson's Chi-squared test
##
## data:  heart_data$Race and heart_data$HeartDisease
## X-squared = 844.31, df = 5, p-value < 2.2e-16
```

```
chisq.test(heart_data$Diabetic, heart_data$HeartDisease)
```

```
##
##  Pearson's Chi-squared test
##
## data:  heart_data$Diabetic and heart_data$HeartDisease
## X-squared = 10960, df = 3, p-value < 2.2e-16
```

```
chisq.test(heart_data$PhysicalActivity, heart_data$HeartDisease)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  heart_data$PhysicalActivity and heart_data$HeartDisease
## X-squared = 3199, df = 1, p-value < 2.2e-16
```

```
chisq.test(heart_data$Asthma, heart_data$HeartDisease)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  heart_data$Asthma and heart_data$HeartDisease
## X-squared = 548.85, df = 1, p-value < 2.2e-16
```

```
chisq.test(heart_data$KidneyDisease, heart_data$HeartDisease)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  heart_data$KidneyDisease and heart_data$HeartDisease
## X-squared = 6739.2, df = 1, p-value < 2.2e-16
```

```
chisq.test(heart_data$SkinCancer, heart_data$HeartDisease)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  heart_data$SkinCancer and heart_data$HeartDisease
## X-squared = 2783.6, df = 1, p-value < 2.2e-16
```

```
# Fit a logistic regression model to identify significant predictors of heart disease
# Converting the hear disease column to 0 and 1
heart_data$HeartDisease <- ifelse(heart_data$HeartDisease == "No", 0, 1)

heart_model <- glm(HeartDisease ~ ., data = heart_data, family = binomial)

#Based on this output, BMI, Smoking(Yes), AlcoholDrinking(Yes), Stroke(Yes),
#PhysicalHealth, MentalHealth, DiffWalking(Yes), Sex(Male), AgeCategory,
#Diabetic(Yes), GenHealthFair, GenHealthGood, GenHealthPoor, GenHealthVeryGood,
#SleepTime, and Asthma(Yes) are all statistically significant predictors of heart disease.


#AgeCategory25-29, DiabeticYes (during pregnancy), PhysicalActivityYes, RaceOther,
#and RaceWhite are not statistically significant predictors of heart disease.

#The z-score is used to assess the statistical significance of each variable in
#the model. The larger the absolute value of the z-score, the further the estimate
#of the coefficient is from zero, indicating a greater impact on the outcome variable

summary(heart_model)
```

```
##
## Call:
## glm(formula = HeartDisease ~ ., family = binomial, data = heart_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1278  -0.4108  -0.2440  -0.1295   3.6087
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -6.2935985  0.1147521 -54.845  < 2e-16 ***
## BMI                              0.0084849  0.0011442   7.415 1.21e-13 ***
## SmokingYes                       0.3563393  0.0143727  24.793  < 2e-16 ***
## AlcoholDrinkingYes              -0.2404857  0.0335303  -7.172 7.38e-13 ***
## StrokeYes                        1.0470172  0.0226326  46.261  < 2e-16 ***
## PhysicalHealth                   0.0031770  0.0008632   3.681 0.000233 ***
## MentalHealth                     0.0047165  0.0008827   5.343 9.12e-08 ***
## DiffWalkingYes                   0.2129965  0.0181484  11.736  < 2e-16 ***
## SexMale                          0.7079573  0.0145679  48.597  < 2e-16 ***
## AgeCategory25-29                 0.1266777  0.1241825   1.020 0.307684
## AgeCategory30-34                 0.4862160  0.1110927   4.377 1.21e-05 ***
## AgeCategory35-39                 0.5946781  0.1063731   5.590 2.26e-08 ***
## AgeCategory40-44                 0.9957684  0.1000696   9.951  < 2e-16 ***
## AgeCategory45-49                 1.3185963  0.0964945  13.665  < 2e-16 ***
## AgeCategory50-54                 1.7273683  0.0931522  18.544  < 2e-16 ***
## AgeCategory55-59                 1.9646188  0.0916901  21.427  < 2e-16 ***
## AgeCategory60-64                 2.2268220  0.0908512  24.511  < 2e-16 ***
## AgeCategory65-69                 2.4688649  0.0905765  27.257  < 2e-16 ***
## AgeCategory70-74                 2.7544487  0.0905120  30.432  < 2e-16 ***
## AgeCategory75-79                 2.9565214  0.0910504  32.471  < 2e-16 ***
## AgeCategory80 or older           3.2142065  0.0907969  35.400  < 2e-16 ***
## RaceAsian                       -0.5366015  0.0841277  -6.378 1.79e-10 ***
## RaceBlack                       -0.3459636  0.0577517  -5.991 2.09e-09 ***
## RaceHispanic                    -0.2549177  0.0588305  -4.333 1.47e-05 ***
## RaceOther                       -0.0611086  0.0639976  -0.955 0.339649
## RaceWhite                       -0.0781639  0.0515923  -1.515 0.129765
## DiabeticNo, borderline diabetes  0.1289554  0.0418233   3.083 0.002047 **
## DiabeticYes                      0.4770059  0.0167216  28.526  < 2e-16 ***
## DiabeticYes (during pregnancy)   0.1234766  0.1050702   1.175 0.239922
## PhysicalActivityYes              0.0198827  0.0160701   1.237 0.215994
## GenHealthFair                    1.5194699  0.0328658  46.232  < 2e-16 ***
## GenHealthGood                    1.0457251  0.0295919  35.338  < 2e-16 ***
## GenHealthPoor                    1.8995671  0.0409498  46.388  < 2e-16 ***
## GenHealthVery good               0.4713771  0.0303711  15.521  < 2e-16 ***
## SleepTime                       -0.0250549  0.0043391  -5.774 7.73e-09 ***
## AsthmaYes                        0.2775421  0.0192197  14.440  < 2e-16 ***
## KidneyDiseaseYes                 0.5684066  0.0244035  23.292  < 2e-16 ***
## SkinCancerYes                    0.1144361  0.0195037   5.867 4.43e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 186906  on 319794  degrees of freedom
```

```
## Residual deviance: 145150  on 319757  degrees of freedom
## AIC: 145226
##
## Number of Fisher Scoring iterations: 7
```

```
################################################################################

# Heart_Data:Health_Metrics

heart_health <- read.csv("heart.csv")

str(heart_health)
```

```
## 'data.frame':    918 obs. of  12 variables:
##  $ Age           : int  40 49 37 48 54 39 45 54 37 48 ...
##  $ Sex           : chr  "M" "F" "M" "F" ...
##  $ ChestPainType : chr  "ATA" "NAP" "ATA" "ASY" ...
##  $ RestingBP     : int  140 160 130 138 150 120 130 110 140 120 ...
##  $ Cholesterol   : int  289 180 283 214 195 339 237 208 207 284 ...
##  $ FastingBS     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RestingECG    : chr  "Normal" "Normal" "ST" "Normal" ...
##  $ MaxHR         : int  172 156 98 108 122 170 170 142 130 120 ...
##  $ ExerciseAngina: chr  "N" "N" "N" "Y" ...
##  $ Oldpeak       : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
##  $ ST_Slope      : chr  "Up" "Flat" "Up" "Flat" ...
##  $ HeartDisease  : int  0 1 0 1 0 0 0 0 1 0 ...
```

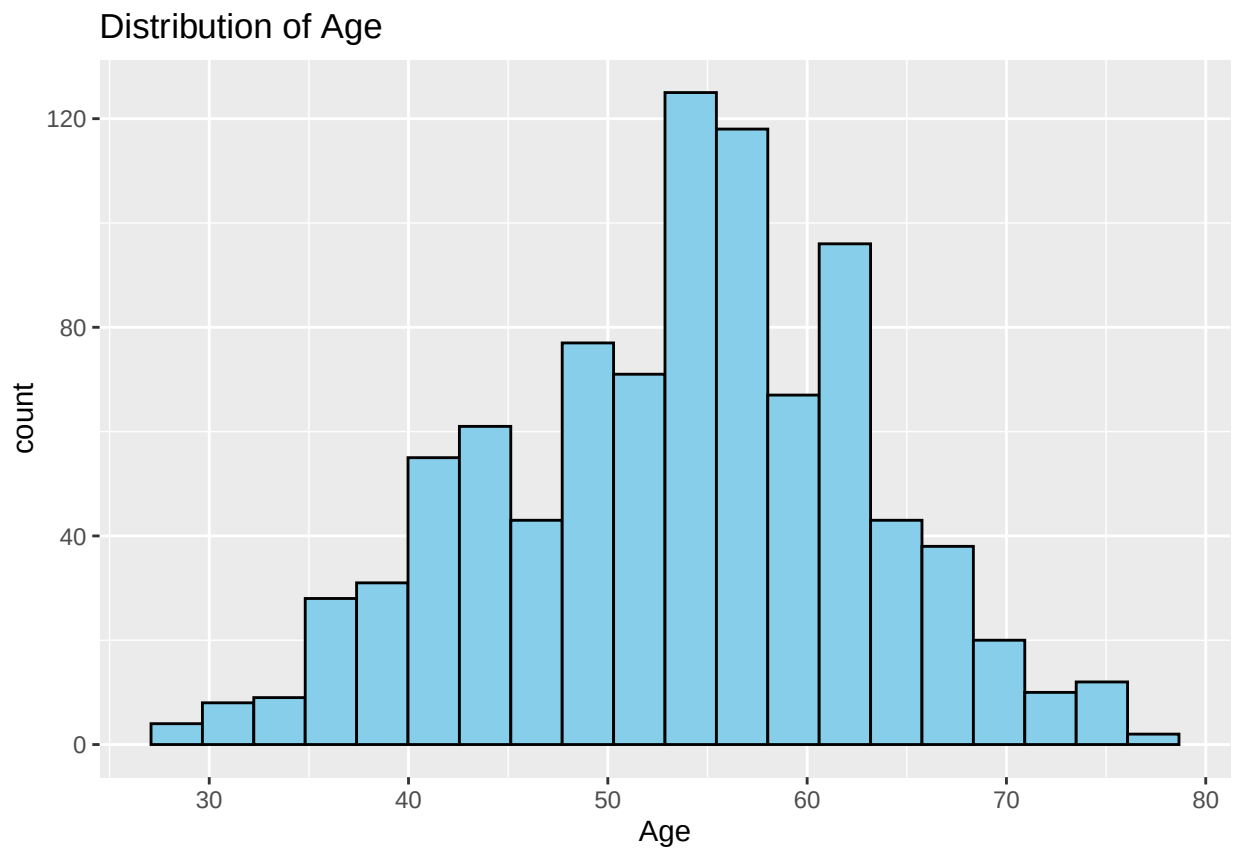
```
summary(heart_health)
```

```
##       Age            Sex            ChestPainType        RestingBP
##  Min.   :28.00   Length:918         Length:918         Min.   :  0.0
##  1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
##  Median :54.00   Mode  :character   Mode  :character   Median :130.0
##  Mean   :53.51                                         Mean   :132.4
##  3rd Qu.:60.00                                         3rd Qu.:140.0
##  Max.   :77.00                                         Max.   :200.0
##   Cholesterol      FastingBS       RestingECG            MaxHR
##  Min.   :  0.0   Min.   :0.0000   Length:918         Min.   : 60.0
##  1st Qu.:173.2   1st Qu.:0.0000   Class :character   1st Qu.:120.0
##  Median :223.0   Median :0.0000   Mode  :character   Median :138.0
##  Mean   :198.8   Mean   :0.2331                      Mean   :136.8
##  3rd Qu.:267.0   3rd Qu.:0.0000                      3rd Qu.:156.0
##  Max.   :603.0   Max.   :1.0000                      Max.   :202.0
##  ExerciseAngina        Oldpeak          ST_Slope          HeartDisease
##  Length:918         Min.   :-2.6000   Length:918         Min.   :0.0000
##  Class :character   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
##  Mode  :character   Median : 0.6000   Mode  :character   Median :1.0000
##                     Mean   : 0.8874                      Mean   :0.5534
##                     3rd Qu.: 1.5000                      3rd Qu.:1.0000
##                     Max.   : 6.2000                      Max.   :1.0000
```

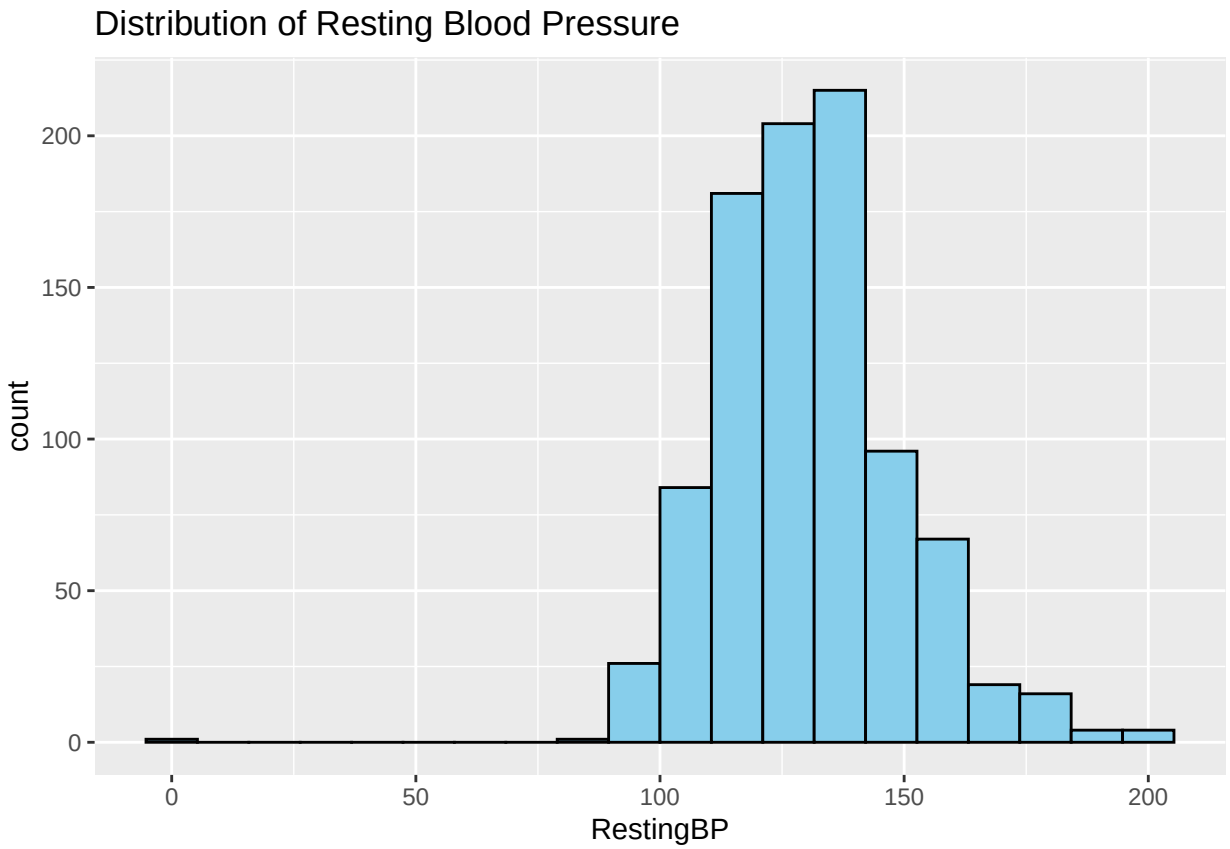
```
sum(is.na(heart_health))
```

```
## [1] 0
```

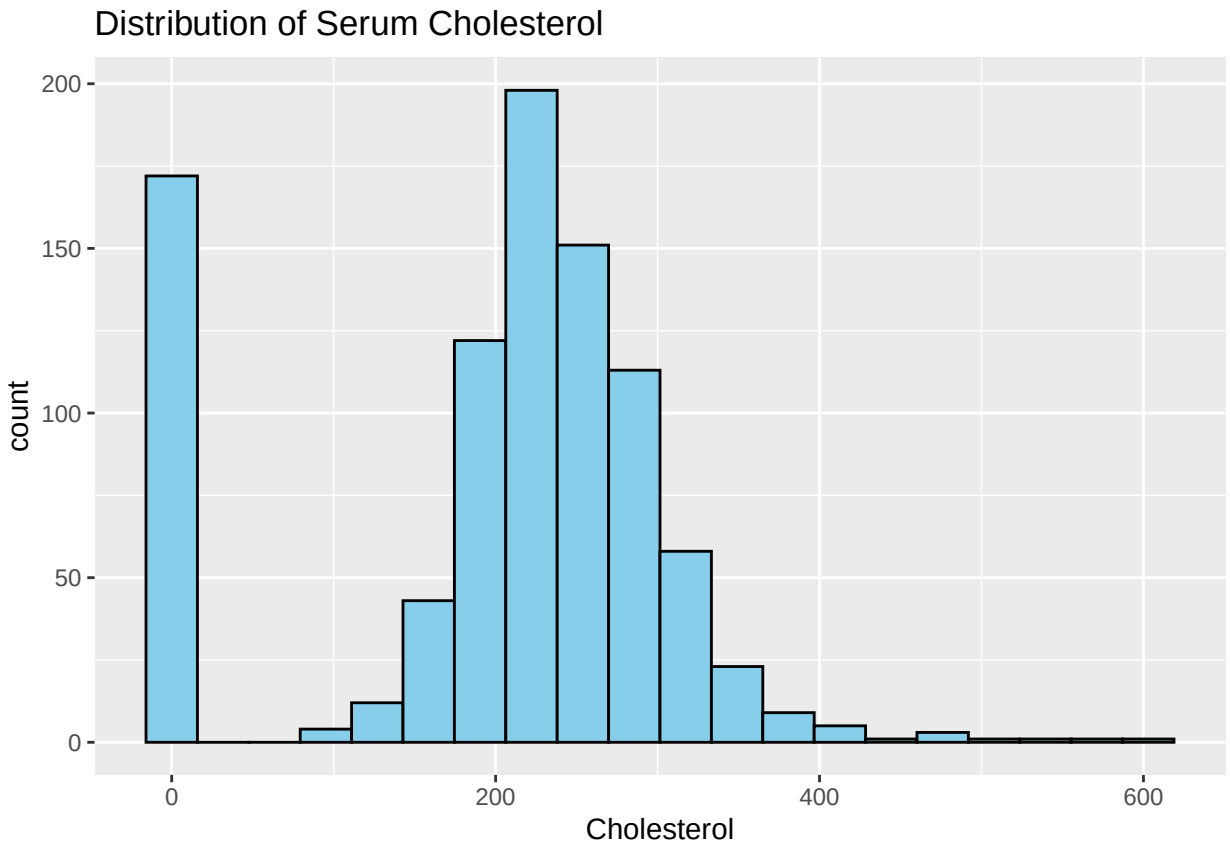
```
ggplot(heart_health, aes(x = Age)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Age")
```



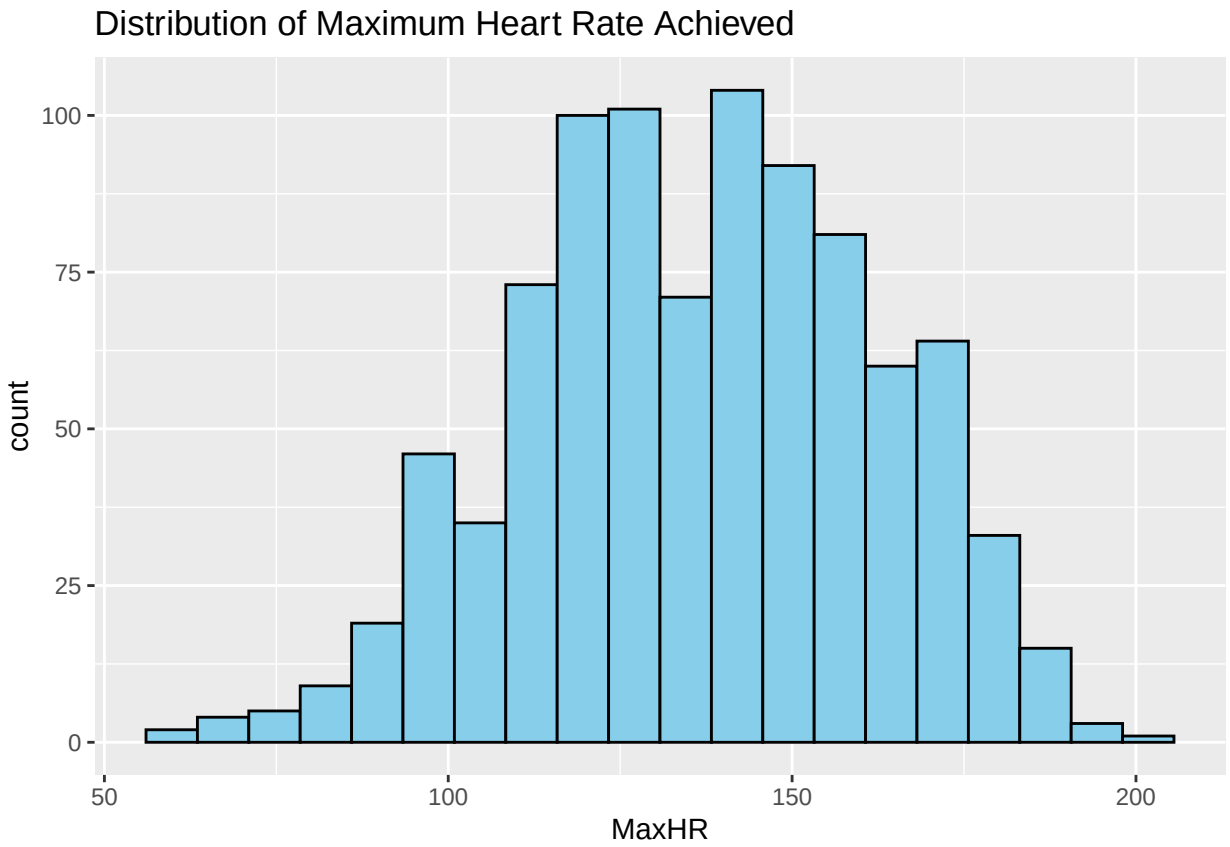

```
ggplot(heart_health, aes(x = RestingBP)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Resting Blood Pressure")
```



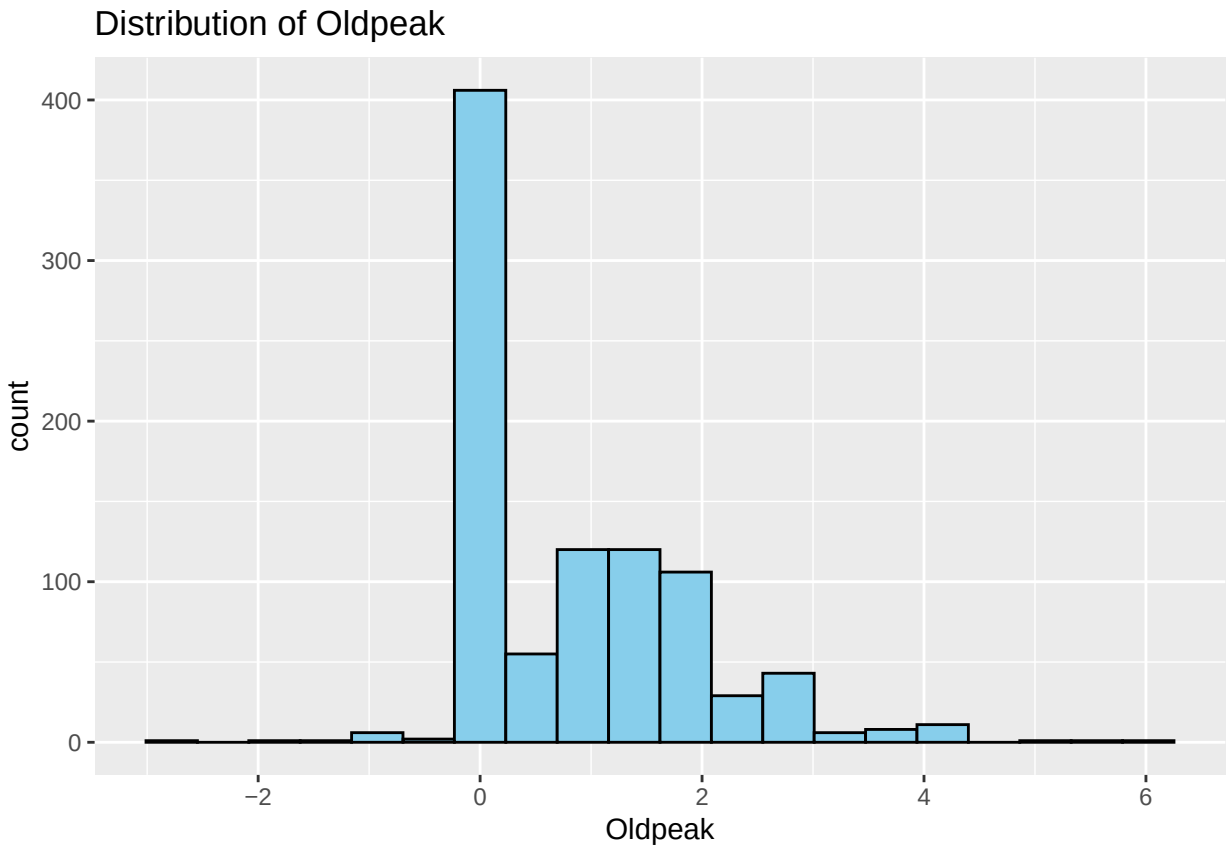

```
ggplot(heart_health, aes(x = Cholesterol)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Serum Cholesterol")
```



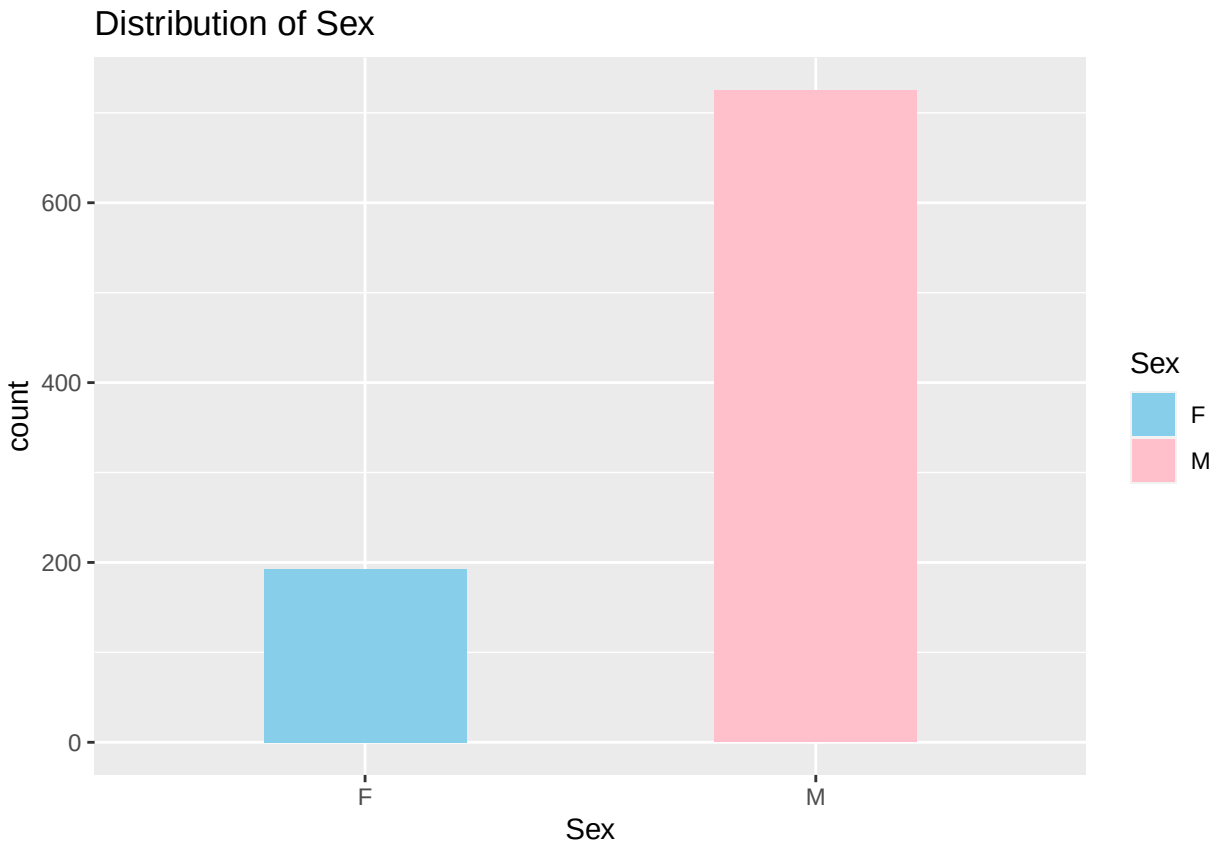

```
ggplot(heart_health, aes(x = MaxHR)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Maximum Heart Rate Achieved")
```



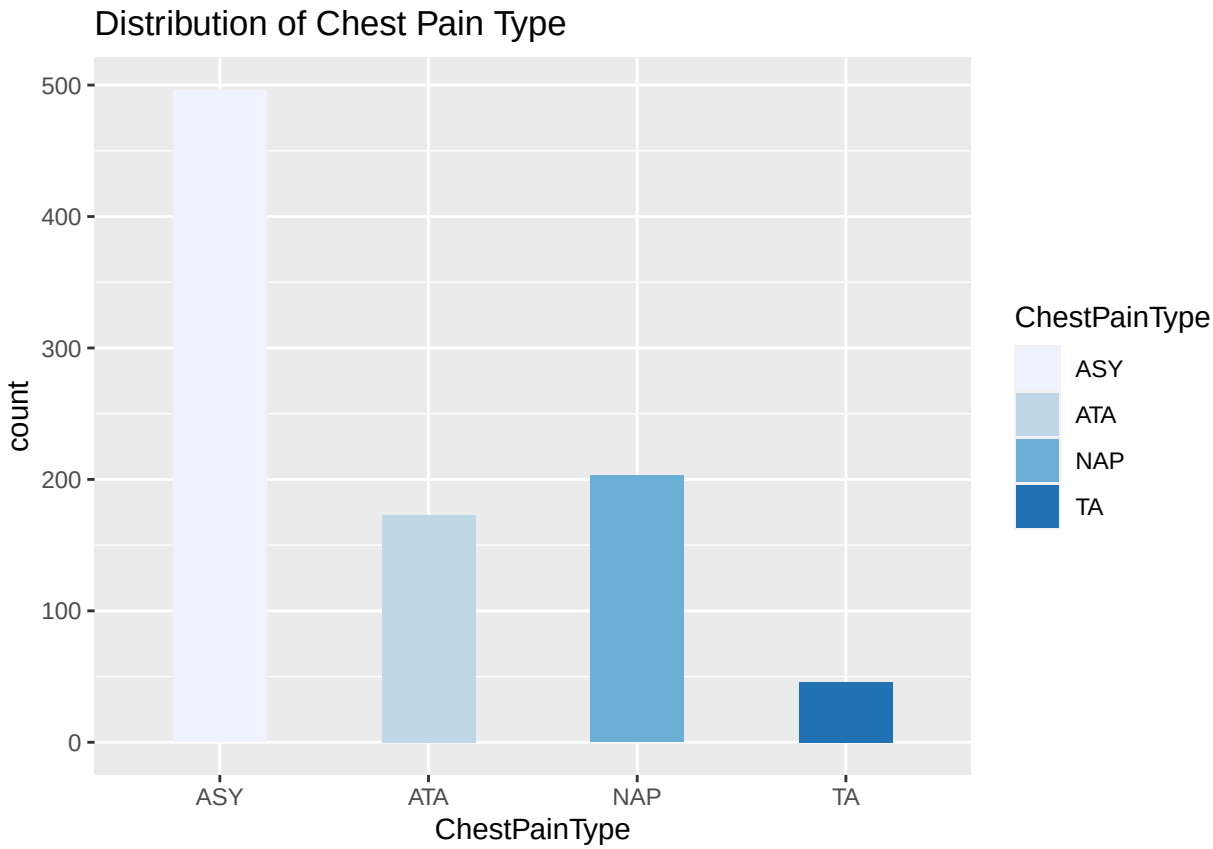

```
ggplot(heart_health, aes(x = Oldpeak)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Oldpeak")
```




```
# Visualize the proportion of categorical variables
ggplot(heart_health, aes(x = Sex, fill = Sex)) +
  geom_bar(position = "dodge", width = 0.45) +
  scale_fill_manual(values = c("skyblue", "pink")) +
  labs(title = "Distribution of Sex")
```

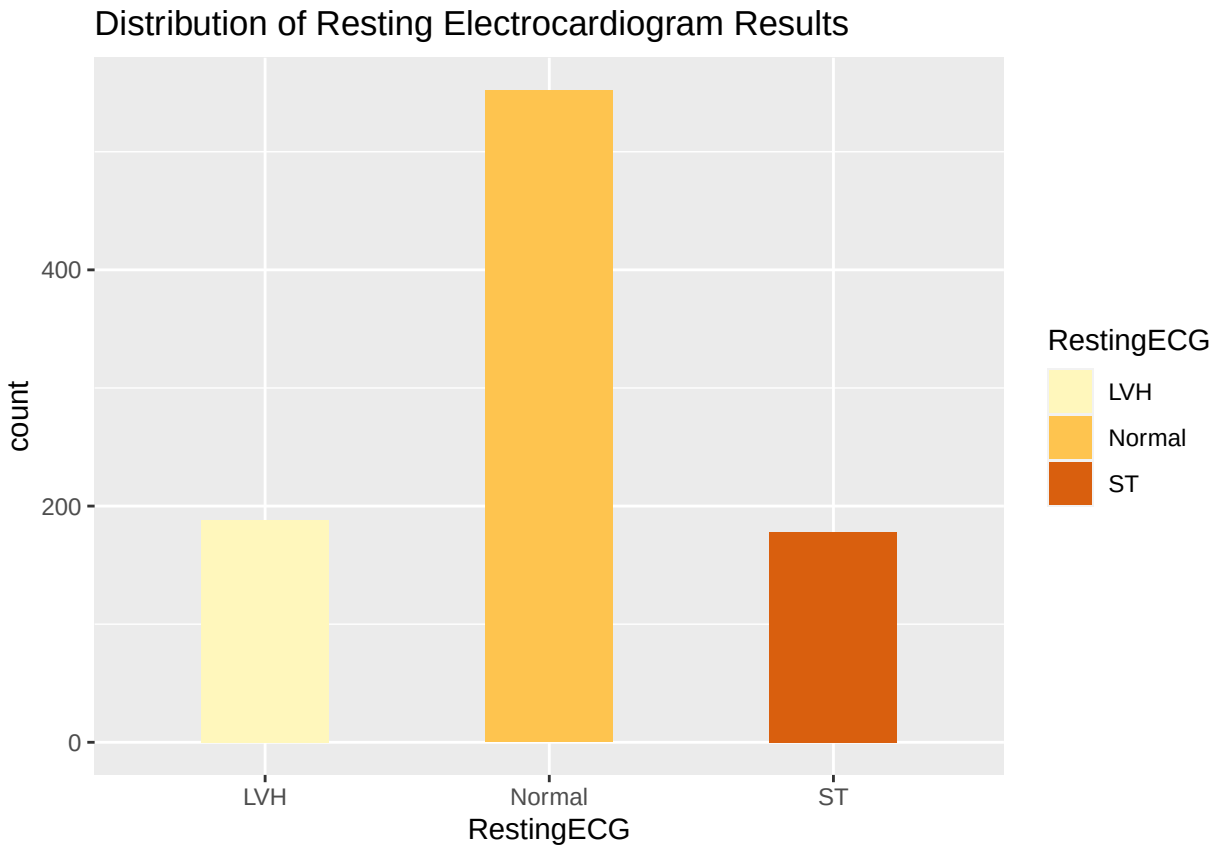



```
ggplot(heart_health, aes(x = ChestPainType, fill = ChestPainType)) +
  geom_bar(position = "dodge", width = 0.45) +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Distribution of Chest Pain Type")
```

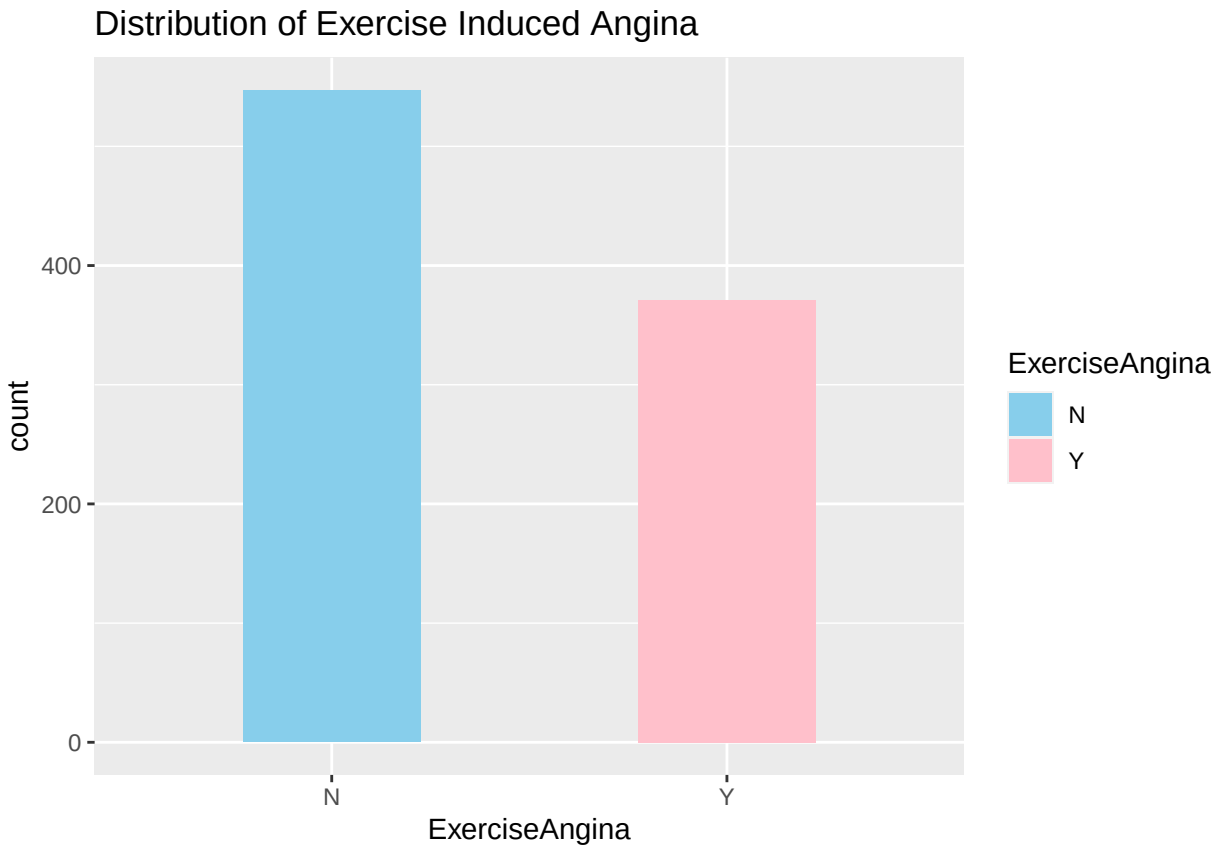



```
ggplot(heart_health, aes(x = FastingBS, fill = FastingBS)) +
  geom_bar(position = "dodge", width = 0.45) +
  scale_fill_manual(values = c("skyblue", "pink")) +
  labs(title = "Distribution of Fasting Blood Sugar")
```

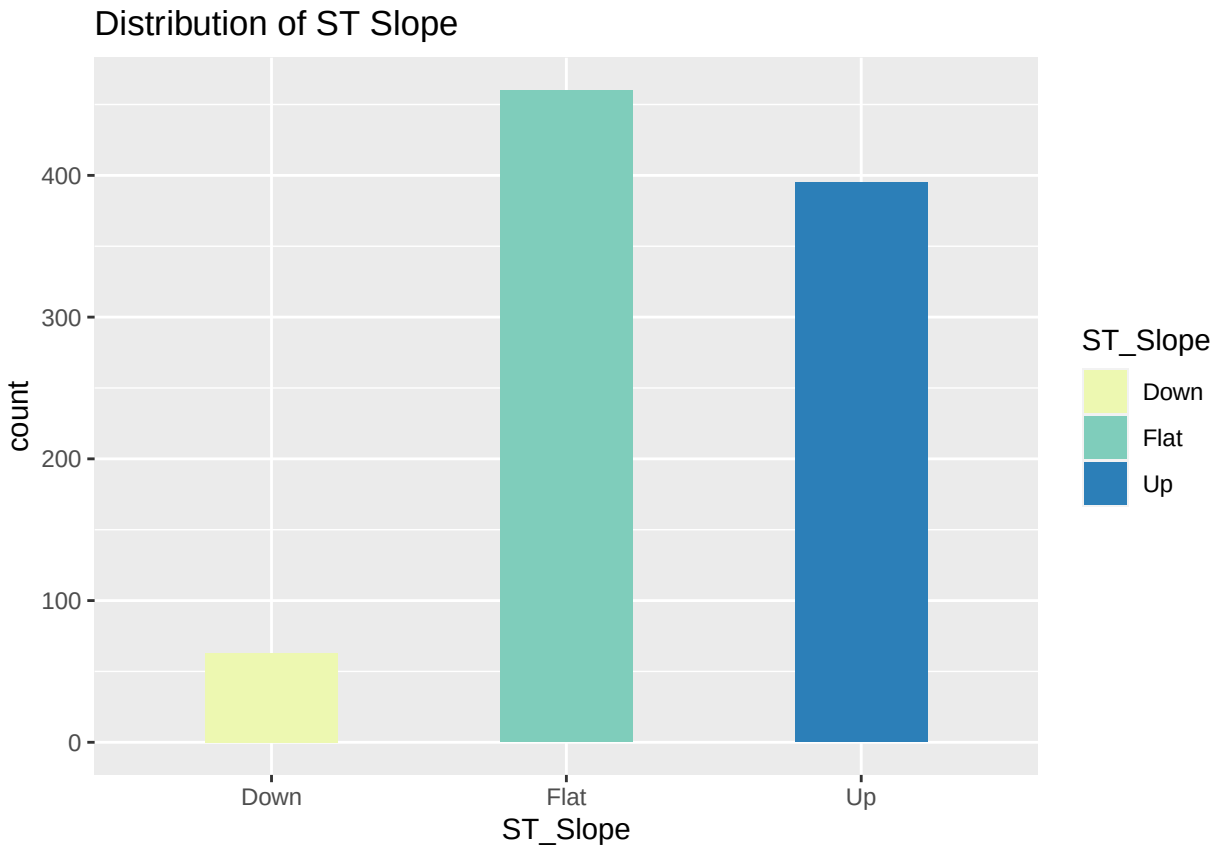

```
ggplot(heart_health, aes(x = RestingECG, fill = RestingECG)) +
  geom_bar(position = "dodge", width = 0.45) +
  scale_fill_brewer(palette = "YlOrBr") +
  labs(title = "Distribution of Resting Electrocardiogram Results")
```




```
ggplot(heart_health, aes(x = ExerciseAngina, fill = ExerciseAngina)) +
  geom_bar(position = "dodge", width = 0.45) +
  scale_fill_manual(values = c("skyblue", "pink")) +
  labs(title = "Distribution of Exercise Induced Angina")
```




```
ggplot(heart_health, aes(x = ST_Slope, fill = ST_Slope)) +
  geom_bar(position = "dodge", width = 0.45) +
  scale_fill_brewer(palette = "YlGnBu") +
  labs(title = "Distribution of ST Slope")
```




```
# performing some statistical tests
#The 95% confidence interval suggests that the true mean age of individuals
#without heart disease is between 4.16 and 6.54 years younger than the mean age
#of individuals with heart disease.
t.test(Age ~ HeartDisease, data = heart_health)
```

```
##
##  Welch Two Sample t-test
##
## data:  Age by HeartDisease
## t = -8.8225, df = 843.69, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -6.538260 -4.158513
## sample estimates:
## mean in group 0 mean in group 1
##        50.55122        55.89961
```

```
#Can reject the null and state there is a significant different between
# the values for patients with heart disease and without heart disease
t.test(RestingBP ~ HeartDisease, data = heart_health)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  RestingBP by HeartDisease
## t = -3.3395, df = 915.14, p-value = 0.0008732
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -6.357955 -1.651148
## sample estimates:
## mean in group 0 mean in group 1
##        130.1805        134.1850
```

```
# even in this we can reject the null hypothesis as the p value is significantly low
t.test(Cholesterol ~ HeartDisease, data = heart_health)
```

```
##
##  Welch Two Sample t-test
##
## data:  Cholesterol by HeartDisease
## t = 7.6269, df = 844.36, p-value = 6.481e-14
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  38.00953 64.35249
## sample estimates:
## mean in group 0 mean in group 1
##        227.1220        175.9409
```

```
# Can reject the null as there is a significant diference in the means for the
# patients with heart disease and with out heart disease
t.test(MaxHR ~ HeartDisease, data = heart_health)
```

```
##
##  Welch Two Sample t-test
##
## data:  MaxHR by HeartDisease
## t = 13.231, df = 877.04, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  17.45551 23.53591
## sample estimates:
## mean in group 0 mean in group 1
##        148.1512        127.6555
```

```
heart_model_1 <- glm(HeartDisease ~ ., data = heart_health, family = binomial)

#Sex, ChestPainTypeATA, ChestPainTypeNAP, ChestPainTypeTA, Cholesterol, FastingBS,
#ExerciseAnginaY, Oldpeak, and ST_SlopeFlat these variables are statistically
#significant in predicting the likelihood of having heart disease
summary(heart_model_1)
```

```
##
## Call:
## glm(formula = HeartDisease ~ ., family = binomial, data = heart_health)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.6531  -0.3747   0.1745   0.4457   2.5778
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.163656   1.416003  -0.822 0.411197
## Age               0.016550   0.013197   1.254 0.209803
## SexM              1.466477   0.279834   5.241 1.60e-07 ***
## ChestPainTypeATA -1.830289   0.326293  -5.609 2.03e-08 ***
## ChestPainTypeNAP -1.685682   0.266001  -6.337 2.34e-10 ***
## ChestPainTypeTA  -1.488392   0.432572  -3.441 0.000580 ***
## RestingBP         0.004194   0.006010   0.698 0.485296
## Cholesterol      -0.004115   0.001087  -3.785 0.000154 ***
## FastingBS         1.136482   0.274999   4.133 3.59e-05 ***
## RestingECGNormal -0.177033   0.271925  -0.651 0.515022
## RestingECGST     -0.268546   0.350020  -0.767 0.442945
## MaxHR            -0.004288   0.005023  -0.854 0.393249
## ExerciseAnginaY   0.900292   0.244513   3.682 0.000231 ***
## Oldpeak           0.380643   0.118466   3.213 0.001313 **
## ST_SlopeFlat      1.453902   0.429086   3.388 0.000703 ***
## ST_SlopeUp       -0.994101   0.450196  -2.208 0.027234 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1262.14  on 917  degrees of freedom
## Residual deviance:  594.19  on 902  degrees of freedom
## AIC: 626.19
##
## Number of Fisher Scoring iterations: 6
```

```
#
# ## Plot one
# heart_data %>%
#   select(AgeCategory, BMI, DiffWalking, GenHealth, MentalHealth, PhysicalHealth, SleepTime, HeartDise
#   mutate(HeartDisease = factor(HeartDisease)) %>%
#   ggplot(aes(x = AgeCategory, y = BMI, color = HeartDisease)) +
#   geom_point() +
#   scale_color_manual(values = c("#800001", "#4B4B4C")) +
#   labs(title = bquote(atop("BMI vs Age Category by Heart Disease", atop(bold(""), ""))),
#        x = "Age category", y = "BMI", color = "Heart disease") +
#   theme(plot.title = element_text(size = 14, face = "bold"),
#         axis.title.x = element_text(size = 12),
#         axis.title.y = element_text(size = 12),
#         axis.text.x = element_text(size = 12, angle = 45, hjust = 1),
#         axis.text.y = element_text(size = 12),
#         legend.title = element_text(size = 12),
#         legend.text = element_text(size = 12),
#         panel.grid.major = element_blank(),
#         panel.grid.minor = element_blank(),
#         panel.border = element_blank(),
#         panel.background = element_blank(),
#         axis.line = element_line(size = .5, color = "black"))
```

```
#
#
#
# ## plot 2
# heart_data %>%
#   group_by(Smoking, HeartDisease) %>%
#   summarise(count = n()) %>%
#   group_by(Smoking) %>%
#   mutate(percent = count / sum(count) * 100,
#          HD_Yes_percent = percent[HeartDisease == 1]) %>%
#   ggplot(aes(x = Smoking, y = percent, fill = as.factor(HeartDisease))) +
#   geom_col(position = "dodge", width = 0.5) +
#   scale_y_continuous(labels = scales::percent_format(scale = 1), limits = c(0, 100), breaks = seq(0,
#   geom_text(aes(label = paste0(round(percent), "%")),
#             position = position_dodge(width = 0.5), size = 4, vjust = -.5,hjust= 0.3) +
#   labs(title = " Impact of Smoking on Heart Disease",
#        x = "Smoking status", y = "Percentage",
#        fill = "Heart disease status") +
#   scale_fill_manual(values = c("#800001", "#4B4B4C"),
#                     labels = c("No", "Yes"),
#                     name = "Heart disease")
#
#
# # Visualization 3: Boxplot of Physical Health by Heart Disease
#
#
# heart_data %>%
#   group_by(HeartDisease) %>%
#   summarise(mean_PhysicalHealth = mean(PhysicalHealth)) %>%
#   ggplot(aes(x = as.factor(HeartDisease), y = mean_PhysicalHealth, fill = as.factor(HeartDisease))) +
#   geom_col() +
#   scale_fill_manual(values = c("#800001", "#4B4B4C"),
#                     labels = c("No", "Yes"),
#                     name = "Heart disease") +
#   labs(title = "Impact of Health Scores on heart disease",
#        x = "Heart disease status", y = "Mean physical health score")
#
#
# # Plot 4
# heart_data %>%
#   mutate(HeartDisease = as.factor(HeartDisease)) %>%
#   group_by(Sex, HeartDisease) %>%
#   summarise(count = n()) %>%
#   mutate(percent = count/sum(count) * 100) %>%
#   ggplot(aes(x = Sex, y = percent, fill = HeartDisease)) +
#   geom_bar(position = "dodge", stat = "identity") +
#   geom_text(aes(label = scales::percent(percent/100),
#                 y = percent+2),
#             position = position_dodge(width = 0.9),
#             size = 3) +
#   ggtitle("Heart Disease by Gender") +
#   scale_fill_manual(values = c("#800001", "#4B4B4C"),
#                     labels = c("No", "Yes"),
```

```
#                     name = "Heart Disease") +
#   labs(x = "Gender", y = "Percentage") +
#   scale_y_continuous(labels = scales::percent_format(scale = 1), limits = c(0, 100))
#
#
# #plot 5
# # library(gridExtra)
# # library(ggpubr)
# #
# # # Subset the data for males and females
# # heart_health_male <- filter(heart_health, Sex == "M")
# # heart_health_female <- filter(heart_health, Sex == "F")
# #
# # # Create a nested pie chart function
# # nested_pie <- function(data, title){
# #   ggpie(data = select(data, count, HeartDisease), values = "count", label = "HeartDisease",
# #         ggtheme = theme_void(), fill = "HeartDisease",
# #         title = title, label_font = 12)
# # }
# #
# # # Create the pie charts for males and females
# # ggarrange(
# #   ggplot(heart_health_male, aes(x = "", fill = ChestPainType)) +
# #     geom_bar(width = 1) +
# #     facet_wrap(~Sex) +
# #     coord_polar(theta = "y") +
# #     theme_void() +
# #     ggtitle("Chest Pain Type by Sex") +
# #     theme(legend.position = "bottom") +
# #     labs(fill = "Chest Pain Type"),
# #
# #   nested_pie(data = heart_health_male %>%
# #                group_by(Sex, ChestPainType, HeartDisease) %>%
# #                summarise(count = n()),
# #              title = "Heart Disease Status"),
# #
# #   ggplot(heart_health_female, aes(x = "", fill = ChestPainType)) +
# #     geom_bar(width = 1) +
# #     facet_wrap(~Sex) +
# #     coord_polar(theta = "y") +
# #     theme_void() +
# #     ggtitle("Chest Pain Type by Sex") +
# #     theme(legend.position = "bottom") +
# #     labs(fill = "Chest Pain Type"),
# #
# #   nested_pie(data = heart_health_female %>%
# #                group_by(Sex, ChestPainType, HeartDisease) %>%
# #                summarise(count = n()),
# #              title = "Heart Disease Status"),
# #
# #   nrow = 2, ncol = 2
# # )
#
```

```
# #plot 5
# # Subset the data for males and females
# heart_health_male <- subset(heart_health, Sex == "M")
# heart_health_female <- subset(heart_health, Sex == "F")
#
# # Define the color palette for heart disease status
# color_palette <- c("#800001", "#4B4B4C")
#
# # Create a grouped bar chart for males
# ggplot(data = heart_health_male, aes(x = ChestPainType, fill = factor(HeartDisease))) +
#   geom_bar(position = position_dodge(), width = .8) +
#   scale_fill_manual(values = color_palette, labels = c("No", "Yes"),
#                     name = "Heart disease") +
#   labs(x = "Chest Pain Type", y = "Count", title = "Heart Disease by Chest Pain Type-Males") +
#   theme_minimal() +
#   theme(plot.title = element_text(size = 14, face = "bold"),
#         axis.title.x = element_text(size = 12),
#         axis.title.y = element_text(size = 12),
#         axis.text.x = element_text(size = 12),
#         axis.text.y = element_text(size = 12),
#         legend.title = element_text(size = 12),
#         legend.text = element_text(size = 12),
#         panel.grid.major = element_blank(),
#         panel.grid.minor = element_blank(),
#         panel.border = element_blank(),
#         panel.background = element_blank(),
#         axis.line = element_line(size = .5, color = "black")) +
#   geom_text(aes(label=..count..), stat='count', position=position_dodge(width=0.8), vjust=-0.5,hjust=
#
# # Create a grouped bar chart for females
# ggplot(data = heart_health_female, aes(x = ChestPainType, fill = factor(HeartDisease))) +
#   geom_bar(position = position_dodge(), width = 0.8) +
#   scale_fill_manual(values = color_palette, labels = c("No", "Yes"),
#                     name = "Heart disease") +
#   labs(x = "Chest Pain Type", y = "Count", title = "Heart Disease by Chest Pain Type-Females") +
#   theme_minimal() +
#   theme(plot.title = element_text(size = 14, face = "bold"),
#         axis.title.x = element_text(size = 12),
#         axis.title.y = element_text(size = 12),
#         axis.text.x = element_text(size = 12),
#         axis.text.y = element_text(size = 12),
#         legend.title = element_text(size = 12),
#         legend.text = element_text(size = 12),
#         panel.grid.major = element_blank(),
#         panel.grid.minor = element_blank(),
#         panel.border = element_blank(),
#         panel.background = element_blank(),
#         axis.line = element_line(size = 0.5, color = "black")) +
#   geom_text(aes(label=..count..), stat='count', position=position_dodge(width=0.8), vjust=-0.5,hjust=
#
```