```
title: "Notebook"
author: "Thao Nguyen"
output: pdf_document
```

```
# Install necessary packages
install.packages('tinytex')
```

```
## Installing package into ’C:/Users/songt/AppData/Local/R/win-library/4.4’
## (as ’lib’ is unspecified)
```

```
## package ’tinytex’ successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\songt\AppData\Local\Temp\RtmpE74Uwu\downloaded_packages
```

```
tinytex::install_tinytex(force = TRUE)
```

```
## tlmgr install tlgpg
```

```
## tlmgr update --self
```

```
## tlmgr install tlgpg
```

```
## tlmgr --repository http://www.preining.info/tlgpg/ install tlgpg
```

```
## tlmgr option repository "https://ctan.math.illinois.edu/systems/texlive/tlnet"
```

```
## tlmgr update --list
```

```
install.packages('ggplot2')
```

```
## Installing package into ’C:/Users/songt/AppData/Local/R/win-library/4.4’
## (as ’lib’ is unspecified)
```

```
## package ’ggplot2’ successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\songt\AppData\Local\Temp\RtmpE74Uwu\downloaded_packages
```

```
install.packages('dplyr')
```

```
## Installing package into ’C:/Users/songt/AppData/Local/R/win-library/4.4’
## (as ’lib’ is unspecified)
```

```
## package ’dplyr’ successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package ’dplyr’
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\songt\AppData\Local\R\win-library\4.4\00LOCK\dplyr\libs\x64\dplyr.dll
## to C:\Users\songt\AppData\Local\R\win-library\4.4\dplyr\libs\x64\dplyr.dll:
## Permission denied
```

```
## Warning: restored ’dplyr’
```

```
##
## The downloaded binary packages are in
##  C:\Users\songt\AppData\Local\Temp\RtmpE74Uwu\downloaded_packages
```

```r
install.packages('tidyr')
```

```
## Installing package into ’C:/Users/songt/AppData/Local/R/win-library/4.4’
## (as ’lib’ is unspecified)
```

```
## package ’tidyr’ successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package ’tidyr’
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\songt\AppData\Local\R\win-library\4.4\00LOCK\tidyr\libs\x64\tidyr.dll
## to C:\Users\songt\AppData\Local\R\win-library\4.4\tidyr\libs\x64\tidyr.dll:
## Permission denied
```

```
## Warning: restored ’tidyr’
```

```
##
## The downloaded binary packages are in
##  C:\Users\songt\AppData\Local\Temp\RtmpE74Uwu\downloaded_packages
```

```r
install.packages('gridExtra')
```

```
## Installing package into ’C:/Users/songt/AppData/Local/R/win-library/4.4’
## (as ’lib’ is unspecified)
```

```
## package ’gridExtra’ successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\songt\AppData\Local\Temp\RtmpE74Uwu\downloaded_packages
```

```r
install.packages('ggExtra')
```

```
## Installing package into ’C:/Users/songt/AppData/Local/R/win-library/4.4’
## (as ’lib’ is unspecified)
```

```
## package ’ggExtra’ successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\songt\AppData\Local\Temp\RtmpE74Uwu\downloaded_packages
```

```r
install.packages('ggridges')
```

```
## Installing package into 'C:/Users/songt/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)

## package 'ggridges' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\songt\AppData\Local\Temp\RtmpE74Uwu\downloaded_packages
```

```r
install.packages('corrplot')
```

```
## Installing package into 'C:/Users/songt/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)

## package 'corrplot' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\songt\AppData\Local\Temp\RtmpE74Uwu\downloaded_packages
```

```r
install.packages('rsample')
```

```
## Installing package into 'C:/Users/songt/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)

## package 'rsample' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\songt\AppData\Local\Temp\RtmpE74Uwu\downloaded_packages
```

```r
# Load the installed packages
library(tidyr)
library(gridExtra)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(ggExtra)
library(ggridges)
library(corrplot)
```

```
## corrplot 0.94 loaded
```

```r
library(rsample)
```

```r
df = read.csv('C:/Users/songt/R projects/Medical Cost Prediction/insurance.csv', header = TRUE)
head(df)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

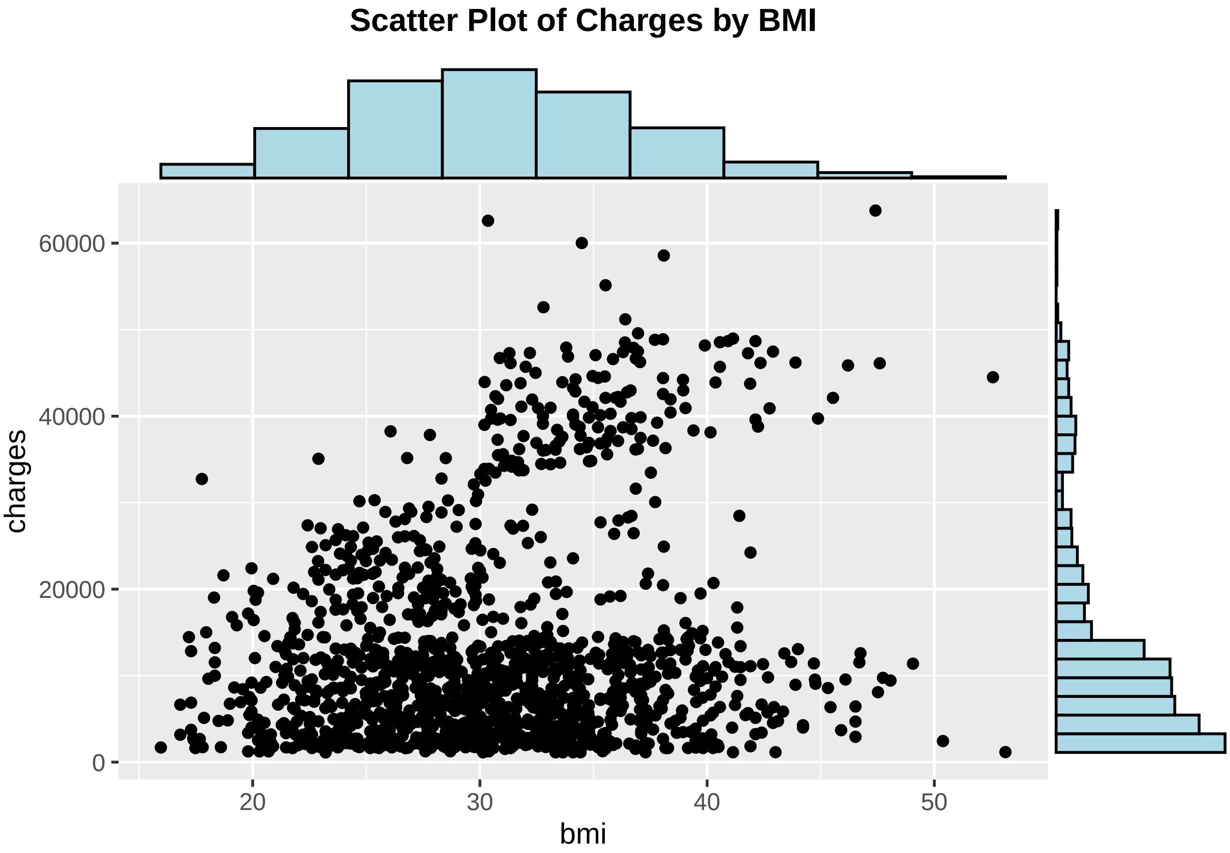
```r
summary(df)
```

```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class  :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode   :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
##     smoker             region             charges
##  Length:1338        Length:1338        Min.   : 1122
##  Class  :character   Class  :character   1st Qu.: 4740
##  Mode   :character   Mode   :character   Median : 9382
##                                          Mean   :13270
##                                          3rd Qu.:16640
##                                          Max.   :63770
```

```r
str(df)
```

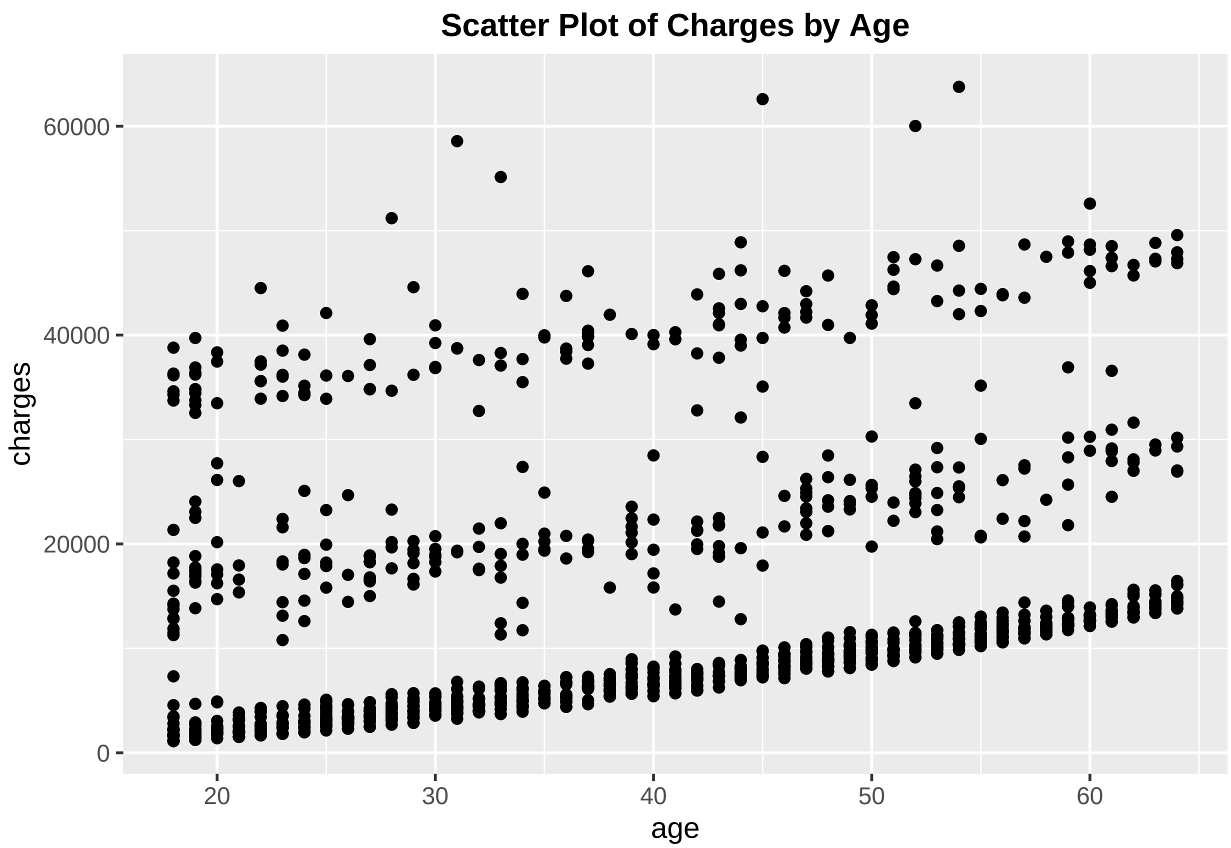
```
## ’data.frame’:    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

```r
# Create the scatter plot of 'charges' versus 'bmi'
g <- ggplot(df, aes(x = bmi, y = charges)) +
  geom_point() +  # Add points to the plot
  theme(legend.position = 'none') +  # Remove legend
  ggtitle("Scatter Plot of Charges by BMI") +  # Add plot title
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5, vjust = 0.5)  # Customize title
  )

# Add marginal histograms to the scatter plot
g1 <- ggMarginal(
  g, type = "histogram", fill = 'lightblue', xparams = list(bins = 10)
)

# Display the final plot
g1
```

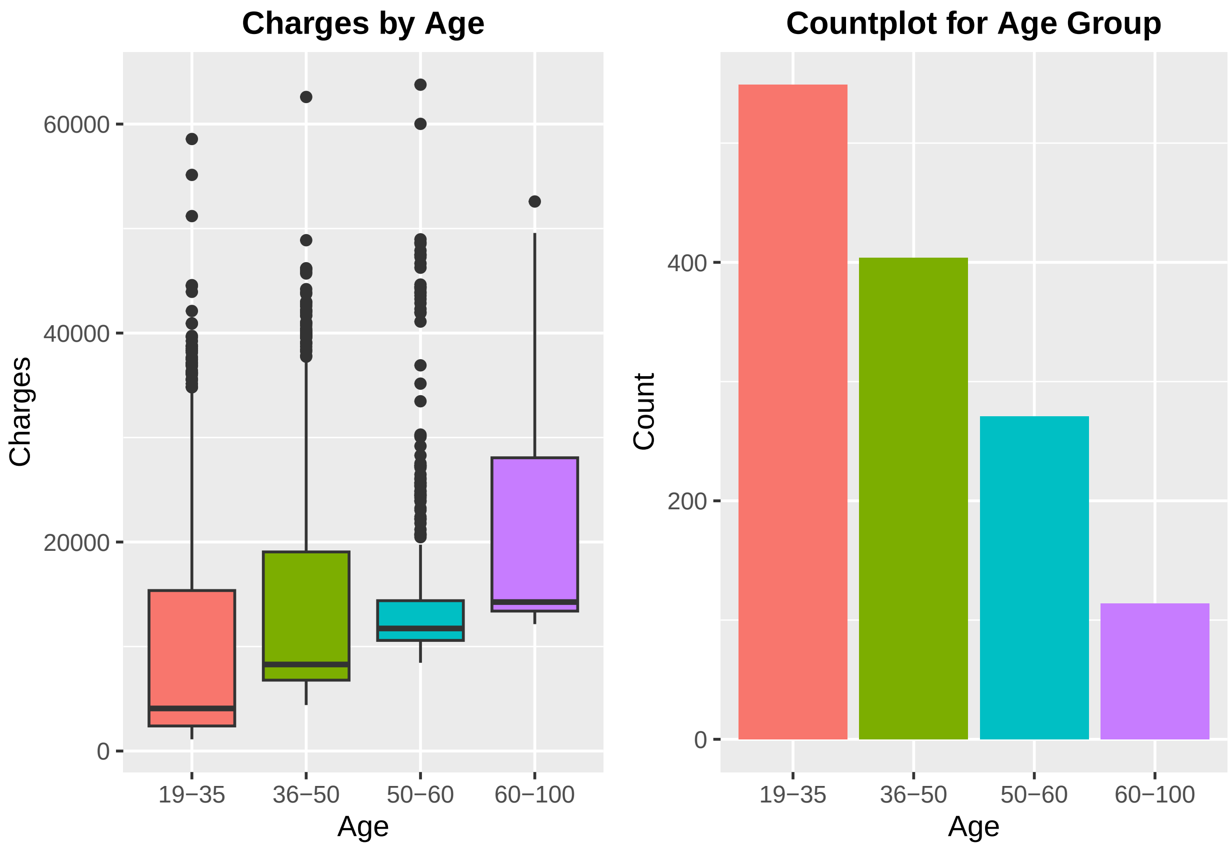


```r
# Create the scatter plot of 'charges' versus 'age'
p <- ggplot(df, aes(x = age, y = charges)) +
  geom_point() +  # Add points to the plot
  theme(legend.position = 'none') +  # Remove legend
  ggtitle("Scatter Plot of Charges by Age") +  # Add plot title
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5, vjust = 0.5)  # Customize title
  )
```

```
# Display the plot
p
```



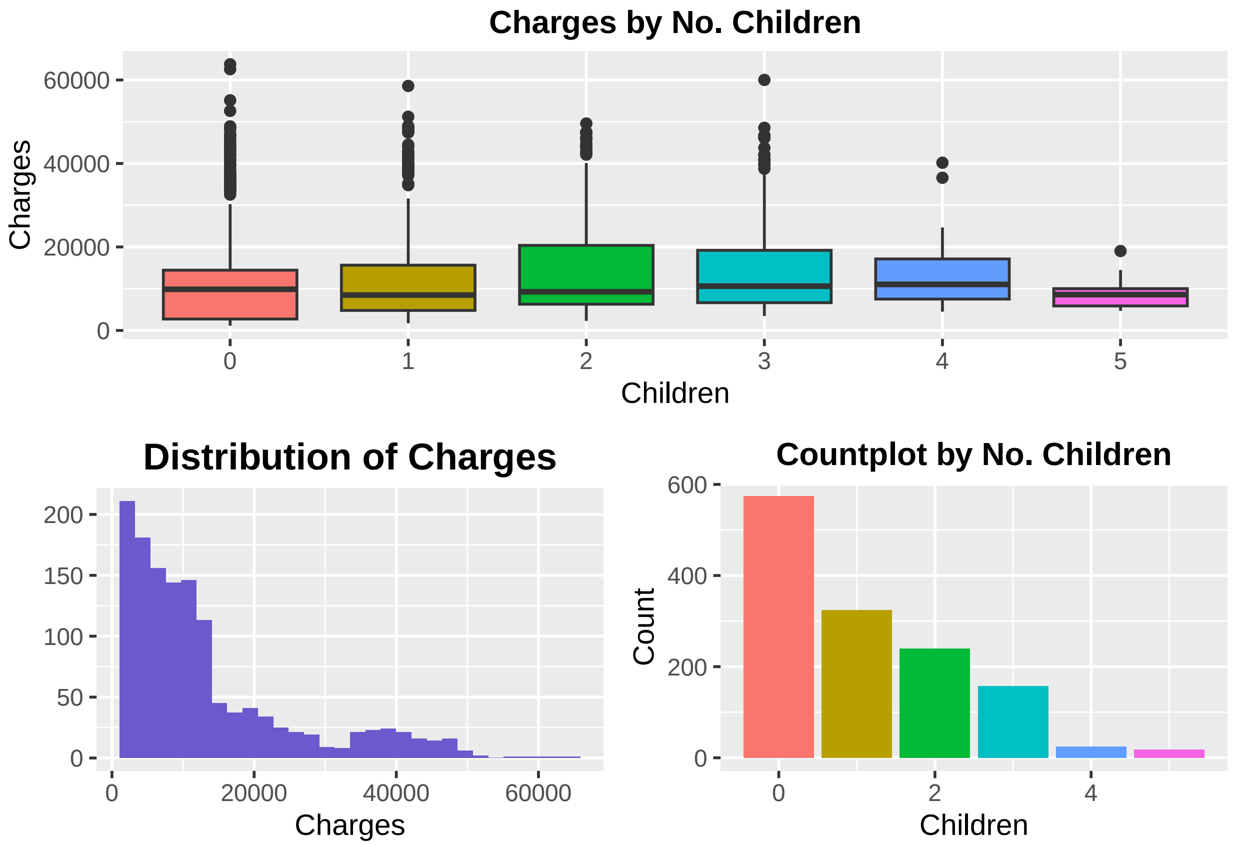
```
# Create age groups in the data frame
df$age_group <- cut(
  df$age,
  breaks = c(0, 18, 35, 50, 60, 100),  # Define age group boundaries
  labels = c('0-18', '19-35', '36-50', '50-60', '60-100'),  # Label each age group
  right = FALSE  # Ensure the interval includes the left value but excludes the right
)

# Boxplot: Charges by Age Group
g1 <- ggplot(df, aes(x = factor(age_group), y = charges, fill = age_group)) +
  geom_boxplot() +  # Create boxplots
  theme(legend.position = 'none') +  # Remove legend
  ggtitle("Charges by Age") +  # Add plot title
  theme(plot.title = element_text(colour = 'black', face = 'bold', size = 12, hjust = 0.5, vjust = 0.5)
  xlab('Age') +  # Label the x-axis
  ylab('Charges')  # Label the y-axis

# Barplot: Count of Age Groups
g2 <- ggplot(df, aes(x = age_group, fill = factor(age_group))) +
  geom_bar() +  # Create a bar plot
  theme(legend.position = 'none') +  # Remove legend
  ggtitle("Countplot for Age Group") +  # Add plot title
  theme(plot.title = element_text(colour = 'black',face = 'bold', size = 12, hjust = 0.5, vjust = 0.5))
  xlab('Age') +  # Label the x-axis
  ylab('Count')  # Label the y-axis

# Arrange the plots side by side
grid.arrange(g1, g2, ncol = 2)
```



```
# Histogram: Distribution of Charges
g1 <- ggplot(df, aes(x = charges)) +
  geom_histogram(fill = 'slateblue') +  # Create a histogram with slate blue color
  theme(legend.position = 'none') +  # Remove legend
  ggtitle("Distribution of Charges") +  # Add plot title
  theme(plot.title = element_text(colour = 'black', face ='bold',size = 14, hjust = 0.5, vjust = 0.5))
  xlab('Charges') +  # Label the x-axis
  ylab('')  # Remove y-axis label

# Boxplot: Charges by Number of Children
g2 <- ggplot(df, aes(x = factor(children), y = charges, fill = factor(children))) +
  geom_boxplot() +  # Create boxplots
  theme(legend.position = 'none') +  # Remove legend
  ggtitle("Charges by No. Children") +  # Add plot title
  theme(plot.title = element_text(face = 'bold', colour = 'black', size = 12, hjust = 0.5, vjust = 0.5)
  xlab('Children') +  # Label the x-axis
  ylab('Charges')  # Label the y-axis
```

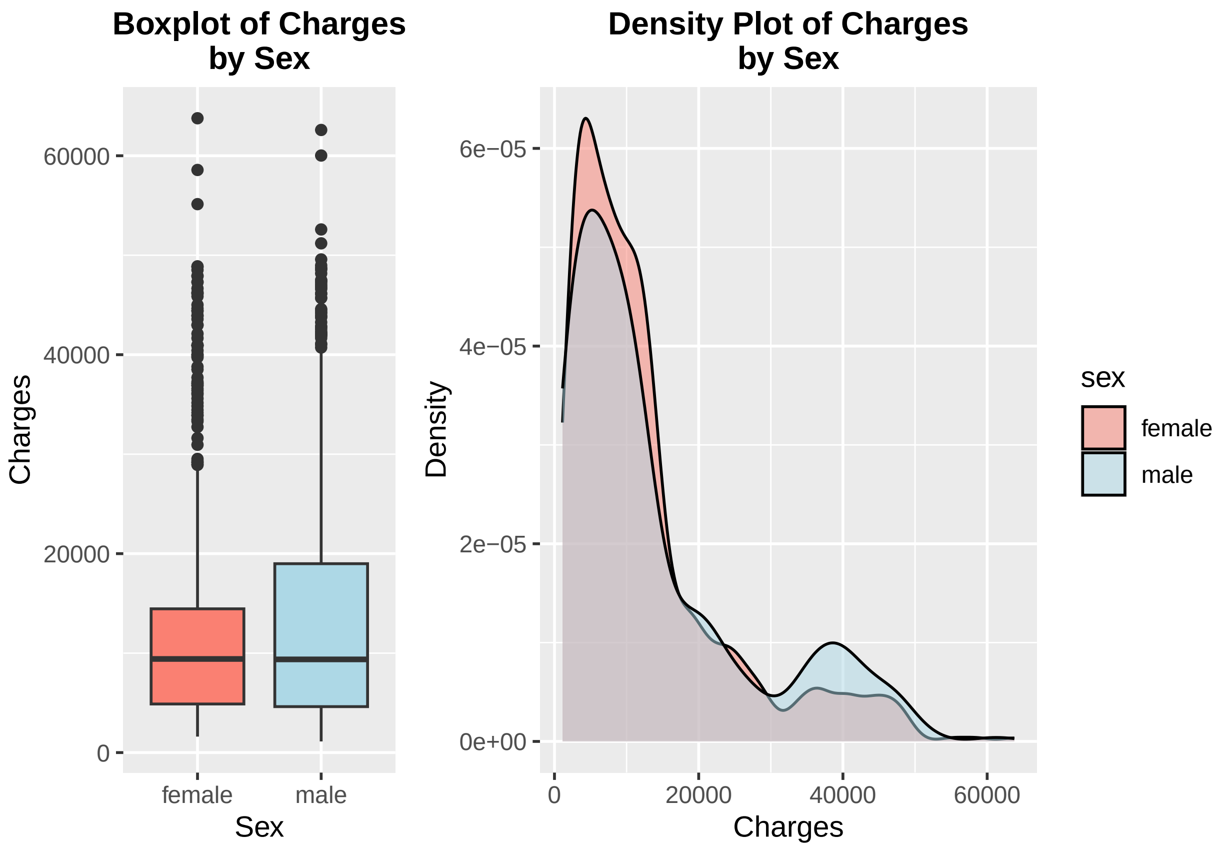
```
# Barplot: Count by Number of Children
g3 <- ggplot(df, aes(x = children, fill = factor(children))) +
  geom_bar() +  # Create a bar plot
  theme(legend.position = 'none') +  # Remove legend
  ggtitle("Countplot by No. Children") +  # Add plot title
  theme(plot.title = element_text(colour = 'black',face ='bold', size = 12, hjust = 0.5, vjust = 0.5)) +
  xlab('Children') +  # Label the x-axis
  ylab('Count')  # Label the y-axis

# Arrange the plots
grid.arrange(
  g2,  # Place the boxplot (g2) on top
  arrangeGrob(g1, g3, ncol = 2),  # Arrange histogram (g1) and bar plot (g3) side by side below g2
  nrow = 2  # Display the plots in two rows
)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

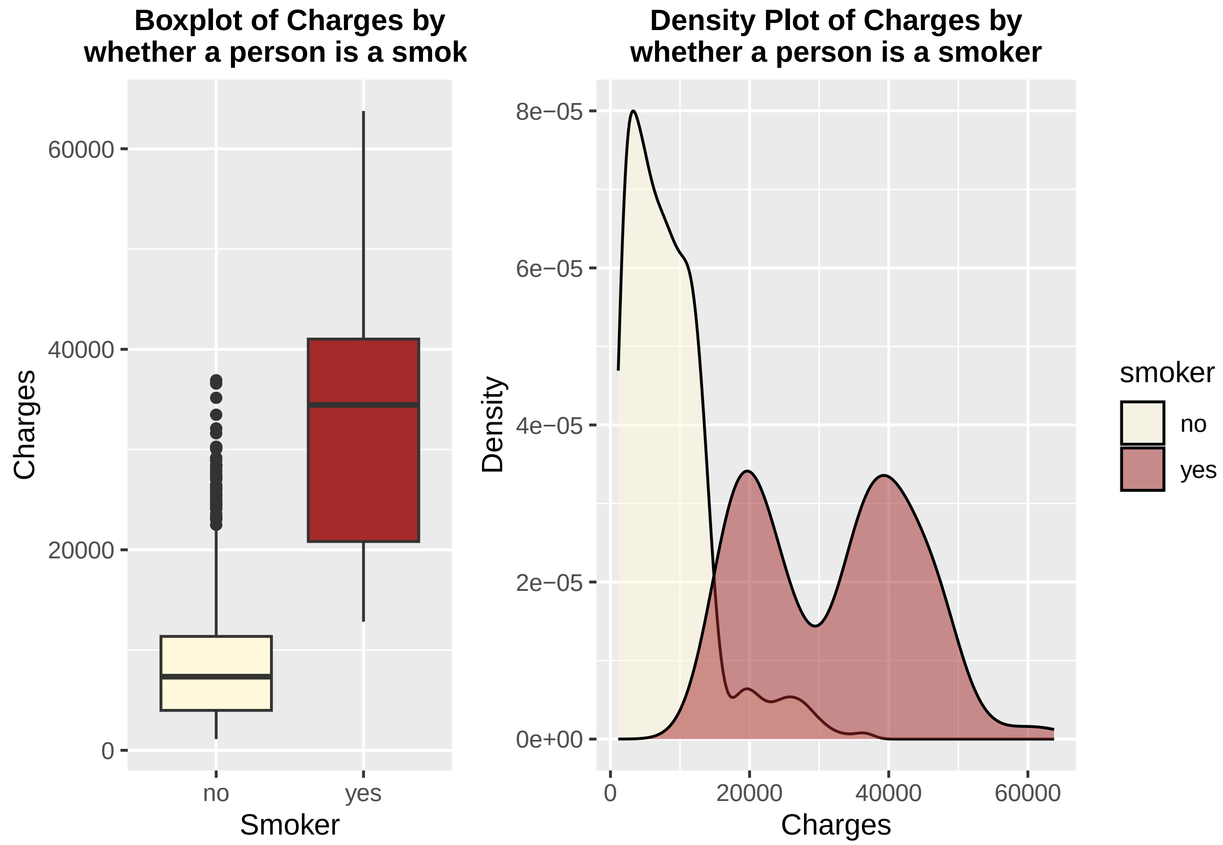


```
# Density Plot: Charges by Sex
g1 <- ggplot(data = df, aes(x = charges, fill = sex)) +
  geom_density(alpha = 0.5) +  # Create a density plot with transparency
  scale_fill_manual(values = c('salmon', 'lightblue')) +  # Set custom colors for sexes
  ggtitle("Density Plot of Charges\nby Sex") +  # Add plot title with a newline for better display
  theme(
    plot.title = element_text(colour = 'black', face ='bold', size = 12, hjust = 0.5, vjust = 0.5)  # C
```

```r
  ) +
  xlab('Charges') +  # Label the x-axis
  ylab('Density')  # Label the y-axis

# Boxplot: Charges by Sex
g2 <- ggplot(data = df, aes(x = factor(sex), y = charges, fill = sex)) +
  geom_boxplot() +  # Create boxplots
  scale_fill_manual(values = c('salmon', 'lightblue')) +  # Set custom colors for sexes
  ggtitle("Boxplot of Charges\nby Sex") +  # Add plot title with a newline for better display
  theme(
    plot.title = element_text(colour = 'black',face ='bold', size = 12, hjust = 0.5, vjust = 0.5),   # C
    legend.position = 'none'  # Remove legend
  ) +
  xlab('Sex') +  # Label the x-axis
  ylab('Charges')  # Label the y-axis

# Arrange the plots side by side
grid.arrange(g2, g1, ncol = 2, widths = c(0.5, 1))
```

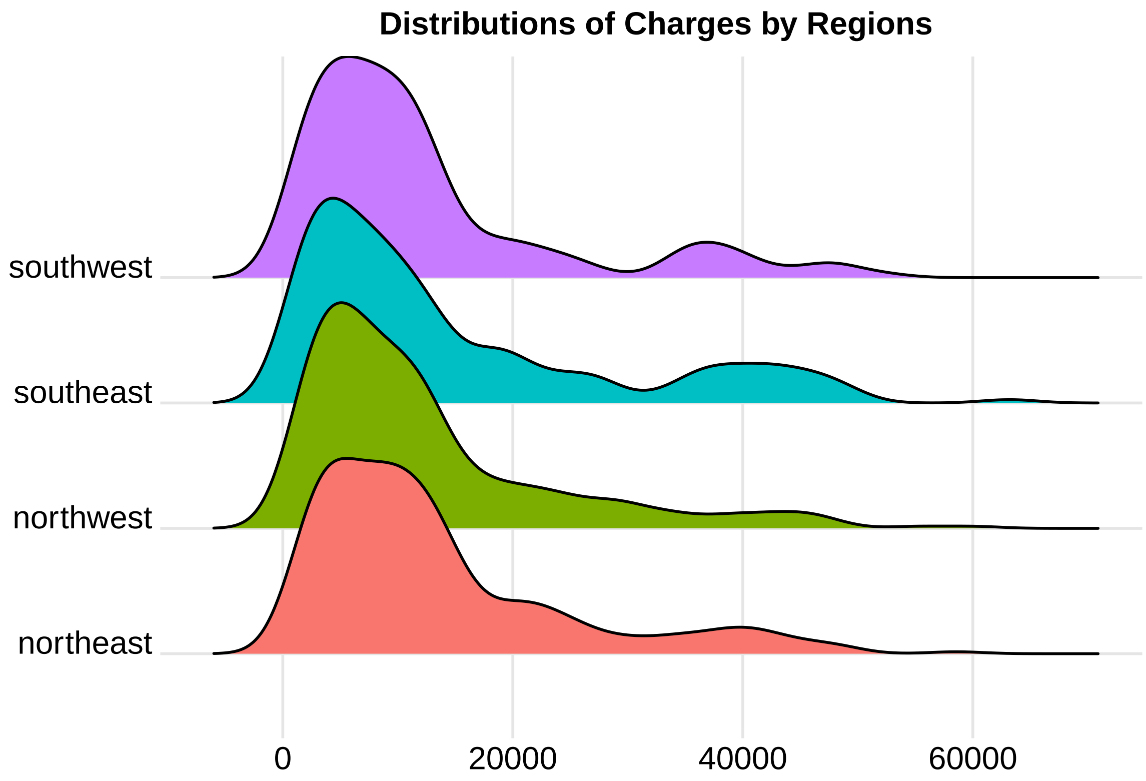


```r
# Density Plot: Charges by Smoking Status
g1 <- ggplot(data = df, aes(x = charges, fill = smoker)) +
  geom_density(alpha = 0.5) +  # Create a density plot
  scale_fill_manual(values = c('cornsilk1', 'brown')) +  # Set custom colors for smokers and non-smoker
  ggtitle("Density Plot of Charges by\nwhether a person is a smoker") +  # Add plot title with a newlin
  theme(
    plot.title = element_text(colour = 'black',face ='bold', size = 11, hjust = 0.5, vjust = 0.5)  # Cu
  ) +
  xlab('Charges') +  # Label the x-axis
  ylab('Density')  # Label the y-axis

# Boxplot: Charges by Smoking Status
g2 <- ggplot(data = df, aes(x = factor(smoker), y = charges, fill = smoker)) +
  geom_boxplot() +  # Create boxplots
  scale_fill_manual(values = c('cornsilk1', 'brown')) +  # Set custom colors for smokers and non-smoker
  ggtitle("Boxplot of Charges by\nwhether a person is a smoker") +  # Add plot title with a newline for
  theme(
    plot.title = element_text(colour = 'black', face ='bold', size = 11, hjust = 0.5, vjust = 0.5),  #
    legend.position = 'none'  # Remove legend
  ) +
  xlab('Smoker') +  # Label the x-axis
  ylab('Charges')  # Label the y-axis

# Arrange the plots in a grid
grid.arrange(g2, g1, ncol = 2, widths = c(0.6, 1))
```

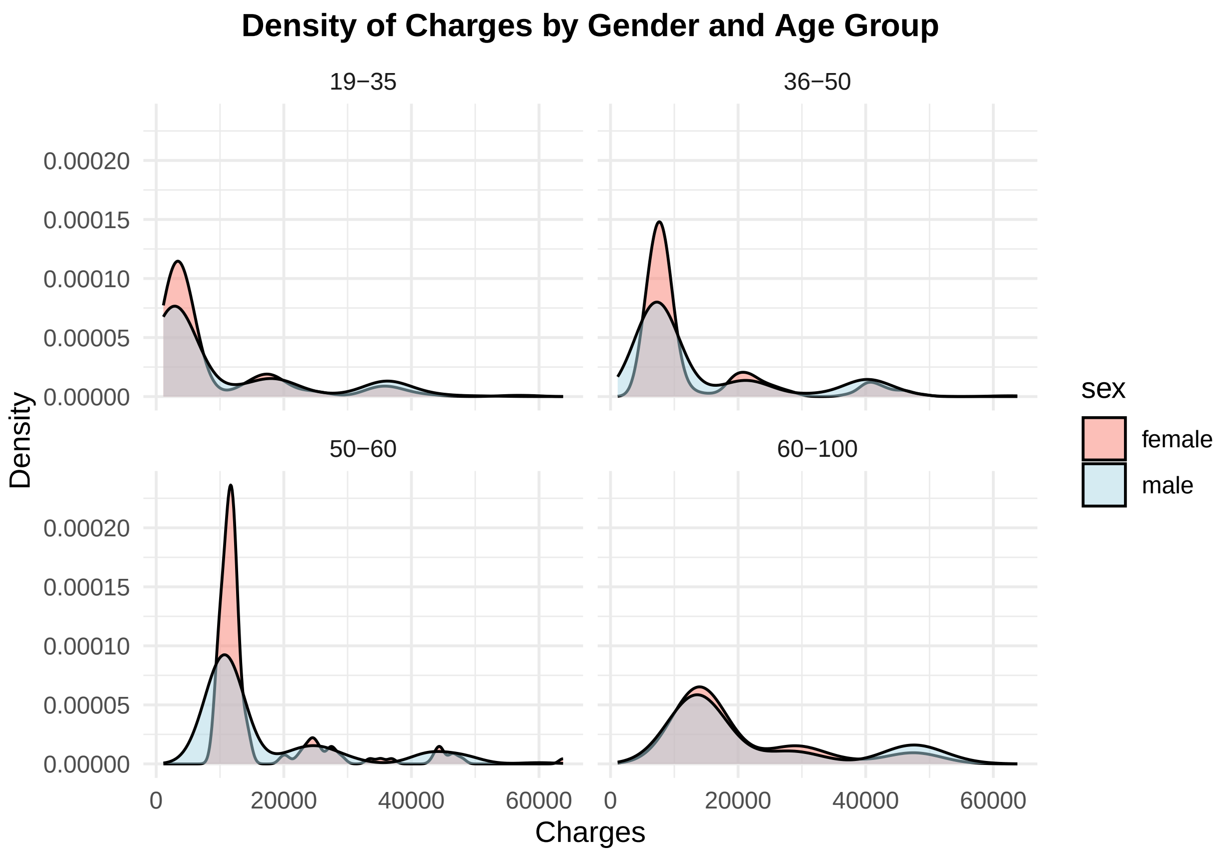


```
# Density Ridges Plot: Distributions of Charges by Region
ggplot(df, aes(x = charges, y = region, fill = region)) +
  geom_density_ridges() +  # Create density ridges to visualize distributions
  theme_ridges() +  # Use the ridges theme for better visual appeal
  theme(
    legend.position = 'none',  # Remove legend
    plot.title = element_text(size = 12, face = 'bold', hjust = 0.5)  # Customize plot title
  ) +
  ggtitle('Distributions of Charges by Regions') +  # Add plot title
  xlab('') +  # No x-axis label
  ylab('')  # No y-axis label
```

```
## Picking joint bandwidth of 2370
```



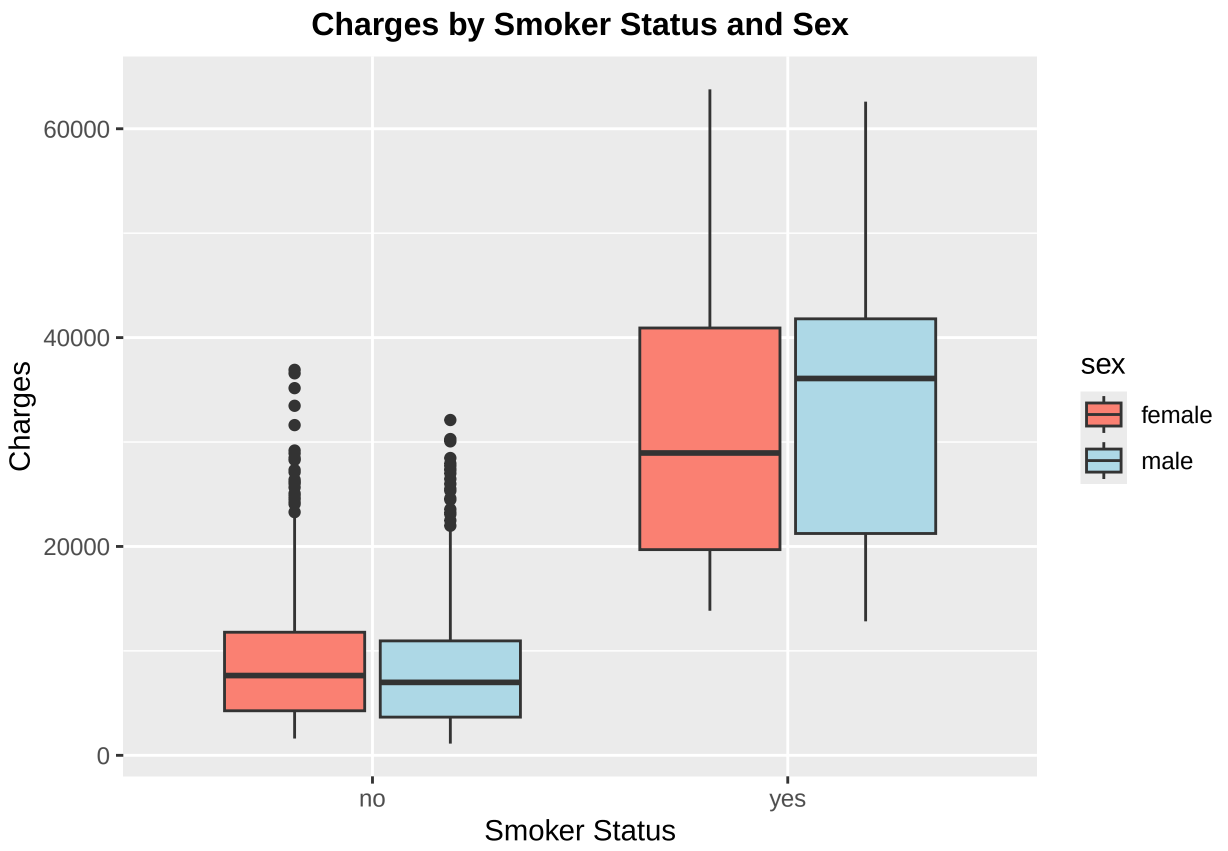
```
# Density Plot: Charges by Gender and Age Group
ggplot(df, aes(x = charges, fill = sex)) +
  geom_density(alpha = 0.5) +  # Create a density plot
  facet_wrap(~age_group) +  # Facet the plot by 'age_group'
  scale_fill_manual(values = c('salmon', 'lightblue')) +  # Set custom colors for genders
  ggtitle("Density of Charges by Gender and Age Group") +  # Add plot title
  xlab('Charges') +  # Label the x-axis
  ylab('Density') +  # Label the y-axis
  theme_minimal() +  # Use a minimal theme for a clean look
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5, size=12)  # Customize plot title font style
  )
```



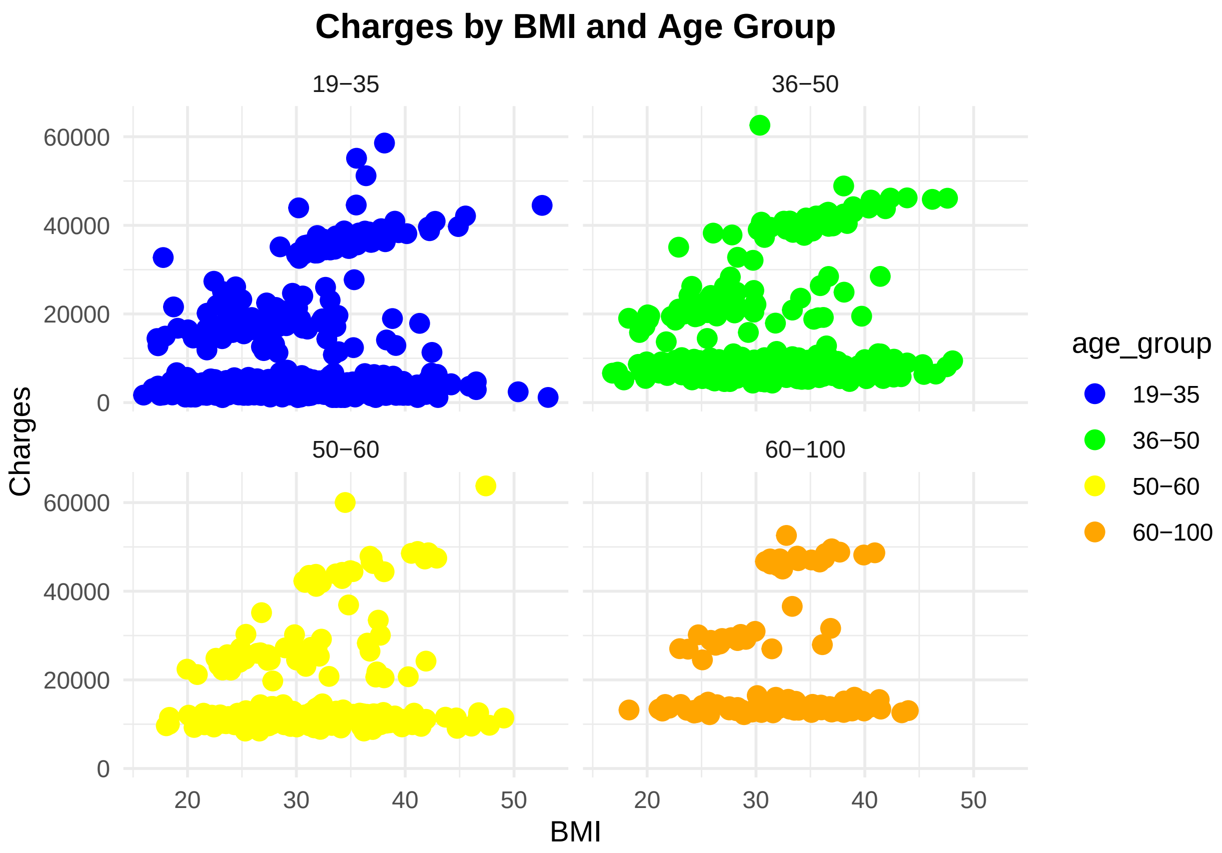
```
# Calculate mean charges by smoker status and gender
charges_by_smoker_gender <- df %>%
  group_by(smoker, sex) %>%
  summarize(mean_charges = mean(charges), .groups = 'drop')  # Calculate mean charges and drop grouping

# Print the summarized data
print(charges_by_smoker_gender)
```

```
## # A tibble: 4 x 3
##   smoker sex    mean_charges
##   <chr>  <chr>         <dbl>
## 1 no     female        8762.
## 2 no     male          8087.
## 3 yes    female       30679.
## 4 yes    male         33042.
```

```
# Boxplot: Charges by Smoker Status and Gender
ggplot(df, aes(x = smoker, y = charges, fill = sex)) +
  geom_boxplot() +  # Create boxplots for charges by smoker status and gender
  scale_fill_manual(values = c('salmon', 'lightblue')) +  # Set custom colors for sexes
  ggtitle("Charges by Smoker Status and Sex") +  # Add plot title
  theme(
    plot.title = element_text(size = 12, face = 'bold', hjust = 0.5)  # Customize title size and alignm
  ) +
  xlab('Smoker Status') +  # Label the x-axis
  ylab('Charges')  # Label the y-axis
```



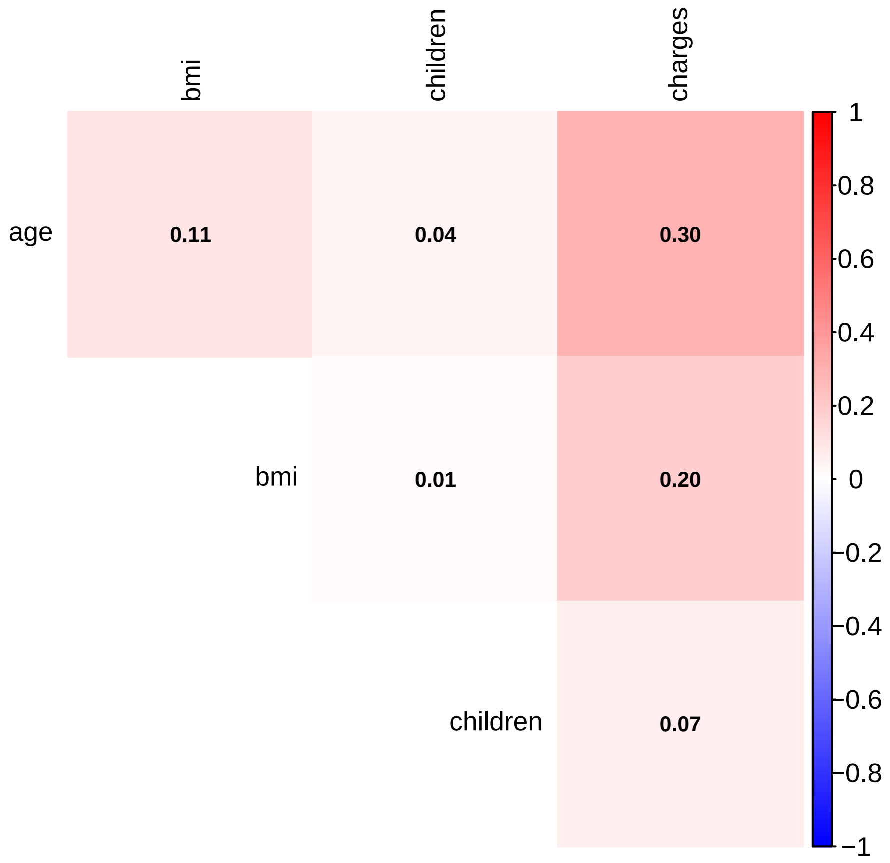
```
# Scatter Plot: Charges by BMI and Age Group
ggplot(df, aes(x = bmi, y = charges, color = age_group)) +
  geom_point(size = 3) +  # Create scatter plot with colored points, size set to 3
  facet_wrap(~age_group) +  # Facet the plot by 'age_group'
  scale_color_manual(values = c('blue', 'green', 'yellow', 'orange')) +  # Set custom colors for age gr
  ggtitle("Charges by BMI and Age Group") +  # Add plot title
  xlab("BMI") +  # Label the x-axis
  ylab("Charges") +  # Label the y-axis
  theme_minimal() +  # Use a minimal theme for a clean look
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5)  # Customize the title font style
  )
```



```r
# Select only numeric columns from the dataframe
df_numeric <- df[sapply(df, is.numeric)]

# Calculate the correlation matrix for numeric columns
corr_matrix <- cor(df_numeric, use = 'pairwise.complete.obs')

# Print the correlation matrix
corr_matrix
```

```
##                age       bmi   children    charges
## age      1.0000000 0.1092719 0.04246900 0.29900819
## bmi      0.1092719 1.0000000 0.01275890 0.19834097
## children 0.0424690 0.0127589 1.00000000 0.06799823
## charges  0.2990082 0.1983410 0.06799823 1.00000000
```

```r
# Visualize the correlation matrix with customized formatting
corrplot(corr_matrix,
         method = 'color',          # Color gradient to represent correlation coefficients
         type = 'upper',            # Show only the upper triangle of the matrix
         tl.cex = 0.8,              # Text label size (0.8 is slightly smaller than default)
         tl.col = 'black',          # Text label color
         number.cex = 0.7,          # Size of the numbers in the plot (0.7 is slightly smaller)
         col = colorRampPalette(c('blue', 'white', 'red'))(200),  # Color gradient from blue to red
         addCoef.col = 'black',     # Color of the correlation coefficients text
         diag = FALSE)              # Exclude the diagonal from the plot
```



```r
# Convert categorical variables to factors
df$sex <- as.factor(df$sex)
df$smoker <- as.factor(df$smoker)
df$region <- as.factor(df$region)
df$age_group <- as.factor(df$age_group)

# Split the dataset into training and testing sets
split <- initial_split(df, prop = 0.8)  # 80% training data, 20% testing data
train_data <- training(split)  # Training data
test_data <- testing(split)    # Testing data

# Fit a linear regression model
model <- lm(charges ~ bmi + age_group + sex + region + smoker, data = train_data)

# Display the summary of the linear regression model
summary(model)
```

```
##
## Call:
## lm(formula = charges ~ bmi + age_group + sex + region + smoker,
##     data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13066.1  -3081.1   -926.8   1725.2  30484.0
##
```

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     -4645.94    1073.64  -4.327 1.65e-05 ***
## bmi               335.13      32.81  10.213  < 2e-16 ***
## age_group36-50   3765.12     455.22   8.271 3.95e-16 ***
## age_group50-60   7566.88     516.16  14.660  < 2e-16 ***
## age_group60-100 10577.73     725.80  14.574  < 2e-16 ***
## sexmale          -202.46     380.71  -0.532   0.5950    
## regionnorthwest  -533.57     544.04  -0.981   0.3269    
## regionsoutheast -1352.17     544.76  -2.482   0.0132 *  
## regionsouthwest  -867.77     543.91  -1.595   0.1109    
## smokeryes       24369.87     479.76  50.796  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 6202 on 1060 degrees of freedom
## Multiple R-squared:  0.7477, Adjusted R-squared:  0.7456 
## F-statistic: 349.1 on 9 and 1060 DF,  p-value: < 2.2e-16
```

```r
# Make predictions on the test data
predictions <- predict(model, newdata = test_data)

# Add the predictions to the test data
test_data$predicted_charges <- predictions

# Display the first few rows of the test data with predictions
head(test_data)
```

```
##   age    sex   bmi children smoker    region   charges age_group
## 1  28   male 33.00        3     no southeast  4449.462     19-35
## 2  32   male 28.88        0     no northwest  3866.855     19-35
## 3  46 female 33.44        1     no southeast  8240.590     36-50
## 4  27   male 42.13        0    yes southeast 39611.758     19-35
## 5  59 female 27.72        3     no southeast 14001.134     50-60
## 6  31   male 36.30        2    yes southwest 38711.000     19-35
##   predicted_charges
## 1          4858.659
## 2          4296.525
## 3          8973.694
## 4         32288.250
## 5         10858.524
## 6         30818.851
```

```r
# Calculate Mean Absolute Error (MAE)
mae <- mean(abs(test_data$charges - test_data$predicted_charges))
mae
```

```
## [1] 4220.262
```

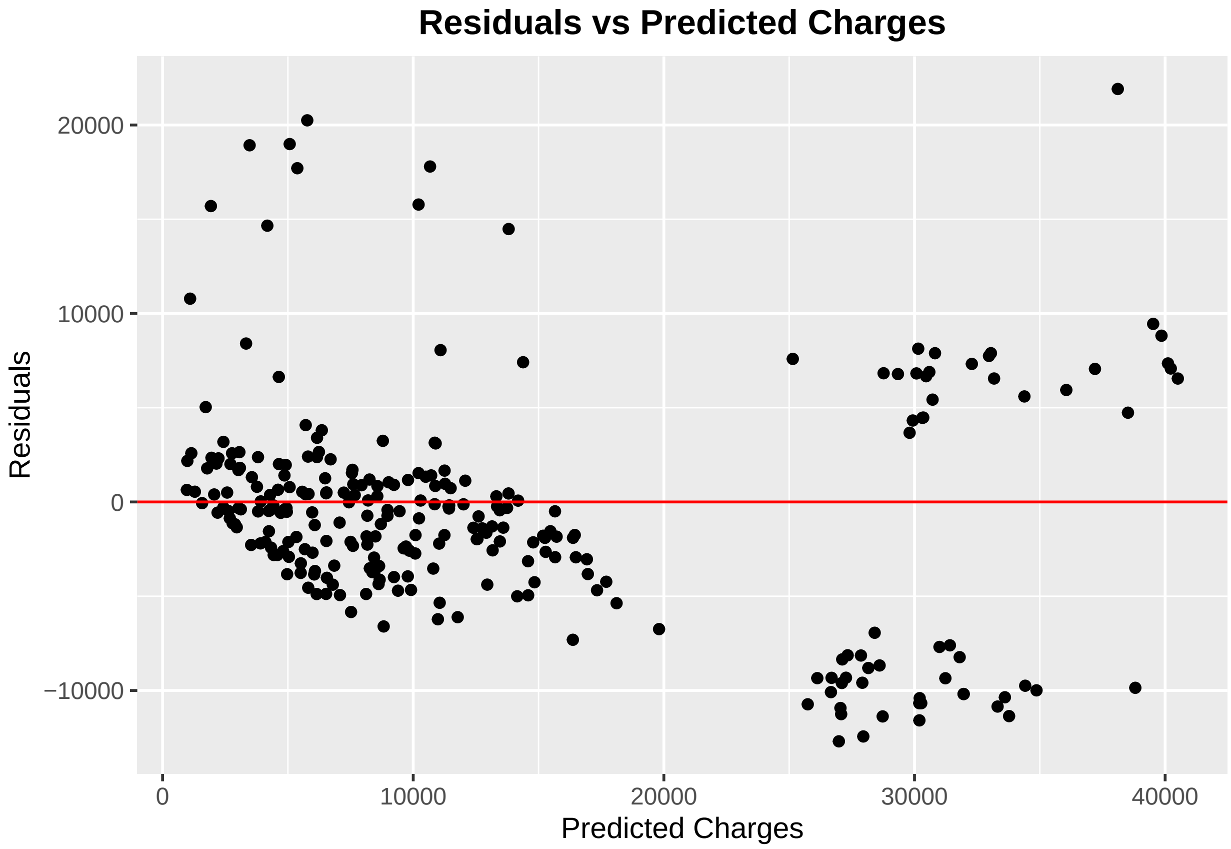
```r
# Calculate Mean Squared Error (MSE)
mse <- mean((test_data$charges - test_data$predicted_charges)^2)
mse
```

```
## [1] 35093505
```

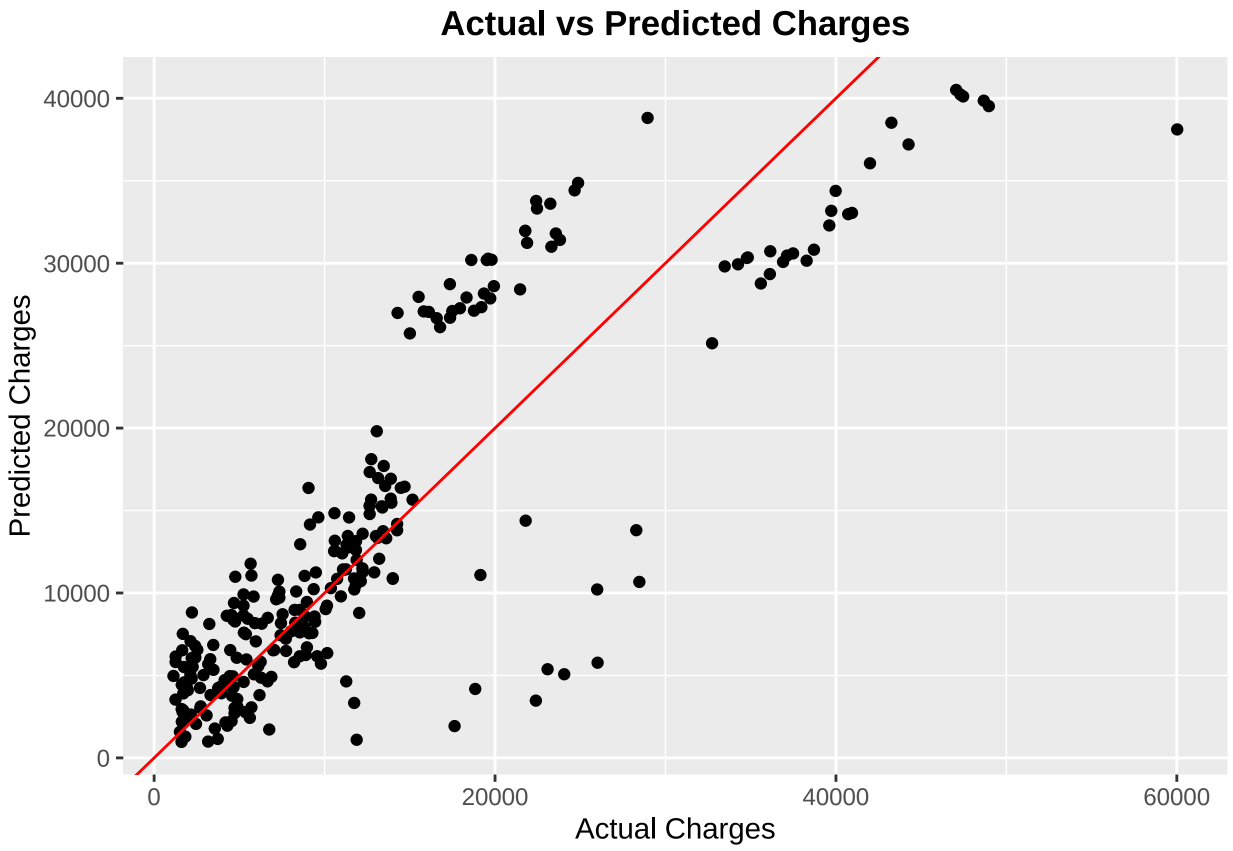
```
# Calculate Root Mean Squared Error (RMSE)
rmse <- sqrt(mse)
rmse
```

```
## [1] 5923.977
```

```
# Plot residuals versus predicted charges
ggplot(test_data, aes(x = predicted_charges, y = charges - predicted_charges)) +
  geom_point() +  # Plot the residuals as points
  geom_hline(yintercept = 0, color = "red") +  # Add a horizontal line at y = 0
  ggtitle("Residuals vs Predicted Charges") +  # Add plot title
  xlab("Predicted Charges") +  # Label x-axis
  ylab("Residuals") + # Label y-axis
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5)  # Customize the title font style
  )
```



```
# Plot actual vs predicted values
ggplot(test_data, aes(x = charges, y = predicted_charges)) +
geom_point() +
geom_abline(intercept = 0, slope = 1, color = "red") +
ggtitle("Actual vs Predicted Charges") +
xlab("Actual Charges") +
ylab("Predicted Charges") +
    # Use a minimal theme for a clean look
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5)  # Customize the title font style
  )
```


Actual vs Predicted Charges
Predicted Charges
Actual Charges