2327510_AnkitRay_MLA1

Ankit Ray

2024-07-14

```r
# Load necessary libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```r
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.3.3
```

```r
library(caret)
```

```
## Loading required package: lattice
```

```r
library(dplyr)
```

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

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

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

```r
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.3.3
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-8
```

```r
# Load the dataset
df <- read.csv("C:/Users/ankit/Desktop/insurance data.csv")

# View the first few rows of the dataset
head(df)
```

```
##   age    sex    bmi children smoker    region   charges  X
## 1  19 female 27.900        0    yes southwest 16884.924 NA
## 2  18   male 33.770        1     no southeast  1725.552 NA
## 3  28   male 33.000        3     no southeast  4449.462 NA
## 4  33   male 22.705        0     no northwest 21984.471 NA
```

```
## 5  32   male 28.880        0     no northwest  3866.855 NA
## 6  31 female 25.740        0     no southeast  3756.622 NA
```

```
# View the structure of the dataset
str(df)
```

```
## 'data.frame':    1338 obs. of  8 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 ...
##  $ X       : logi  NA NA NA NA NA NA ...
```

```
# Summary statistics for the dataset
summary(df)
```

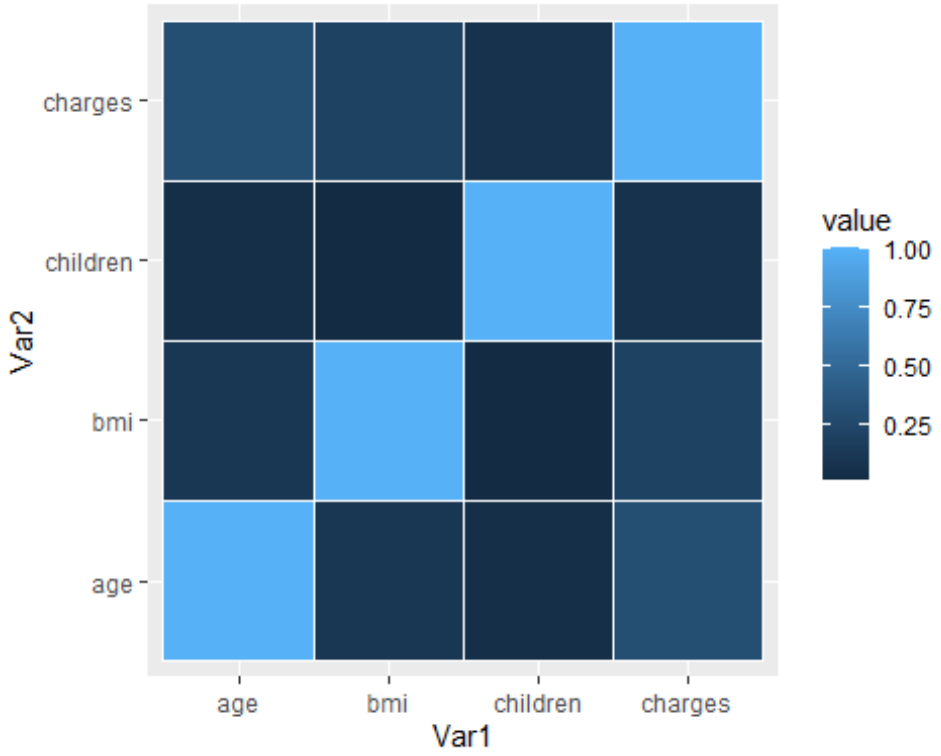
```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      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         X
##  Length:1338        Length:1338        Min.   : 1122   Mode:logical
##  Class :character   Class :character   1st Qu.: 4740   NA's:1338
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

```
# Check for missing values
missing_values <- colSums(is.na(df))
print(missing_values)
```

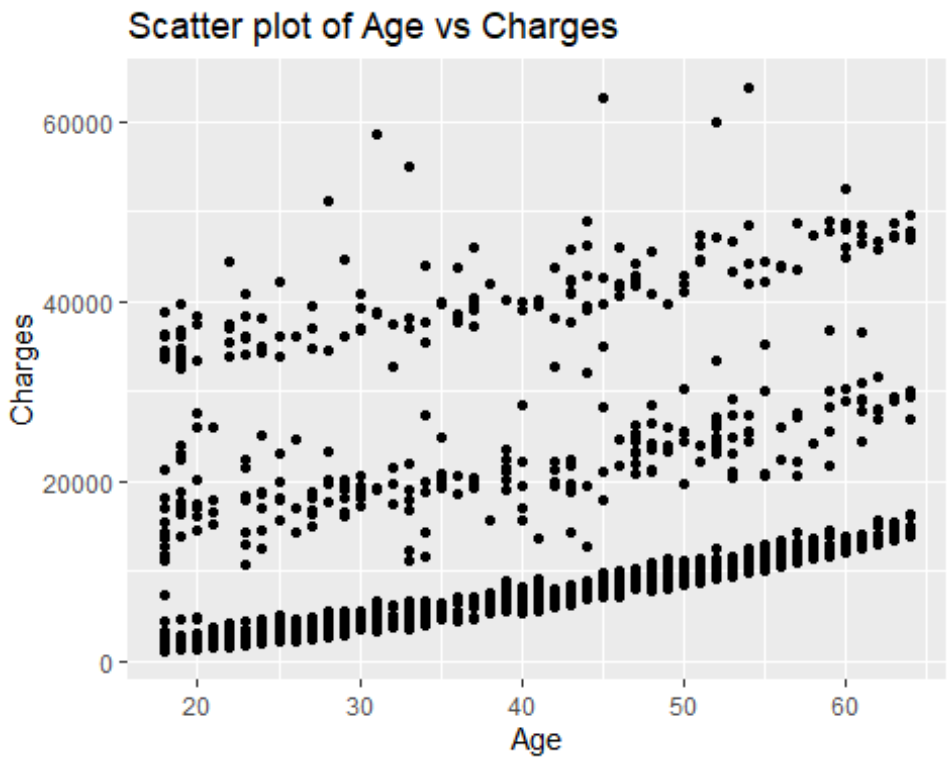
```
##      age      sex      bmi children   smoker   region  charges        X
##        0        0        0        0        0        0        0     1338
```

```
# Select numeric columns for correlation matrix
numeric_columns <- df[, c("age", "bmi", "children", "charges")]

# Create the correlation matrix
correlation_matrix <- cor(numeric_columns)

# Melt the correlation matrix for ggplot2
melted_correlation <- melt(correlation_matrix)

# Create the heatmap
```

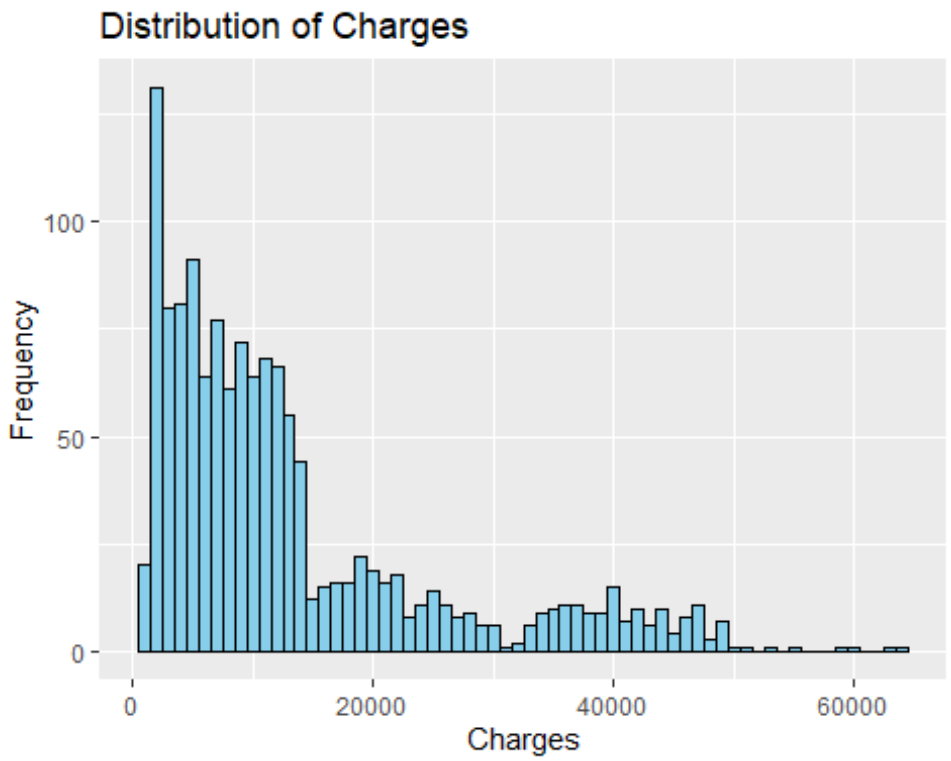
```r
heatmap <- ggplot(data = melted_correlation, aes(x = Var1, y = Var2, fill =
value)) +
  geom_tile(color = "white")

# Show the heatmap
print(heatmap)
```



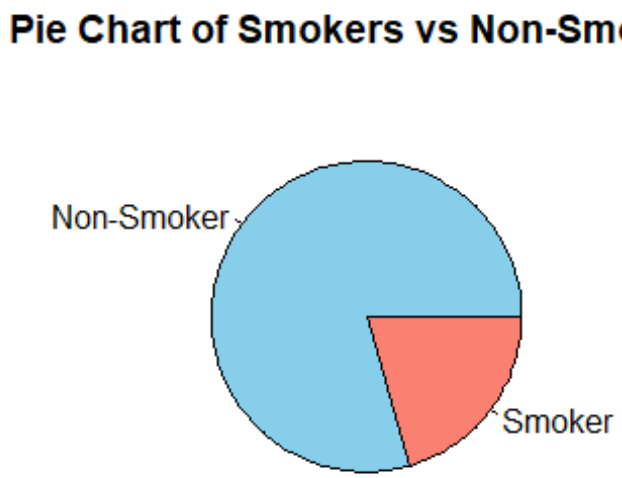

```r
# Scatter plot of Age vs Charges
ggplot(df, aes(x = age, y = charges)) +
  geom_point() +
  labs(x = "Age", y = "Charges", title = "Scatter plot of Age vs Charges")
```



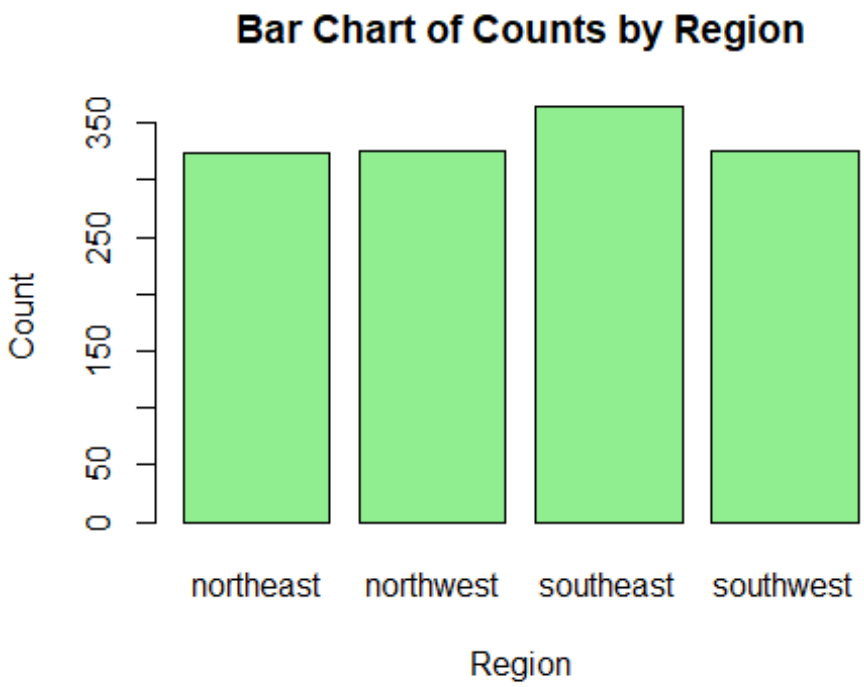

```r
# Histogram of Charges
ggplot(df, aes(x = charges)) +
  geom_histogram(binwidth = 1000, fill = "skyblue", color = "black") +
  labs(x = "Charges", y = "Frequency", title = "Distribution of Charges")
```




```r
# Pie chart for Smoker
smoker_counts <- table(df$smoker)
pie(smoker_counts, labels = c("Non-Smoker", "Smoker"),
    col = c("skyblue", "salmon"),
    main = "Pie Chart of Smokers vs Non-Smokers")
```

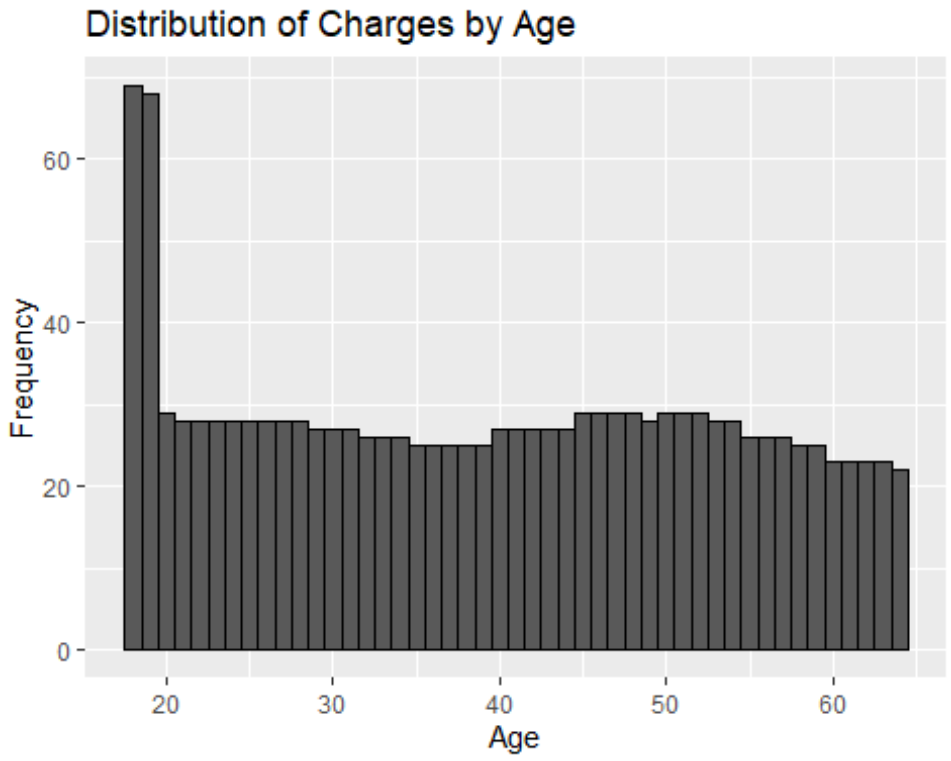



```r
# Bar chart for Region
region_counts <- table(df$region)
barplot(region_counts, col = "lightgreen",
        main = "Bar Chart of Counts by Region",
        xlab = "Region", ylab = "Count")
```




```r
# Histogram of Charges by Age
ggplot(df, aes(x = age, fill = charges)) +
  geom_histogram(binwidth = 1, color = "black") +
  labs(x = "Age", y = "Frequency", title = "Distribution of Charges by Age")
```

```
## Warning: The following aesthetics were dropped during statistical
transformation: fill.
## ℹ This can happen when ggplot fails to infer the correct grouping
structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```r
# MLR

# Convert categorical variables to factors if needed
df <- df %>%
  mutate(
    sex = as.factor(sex),
    smoker = as.factor(smoker),
    region = as.factor(region)
  )

# Check the structure of your data
str(df)
```

```
## 'data.frame':    1338 obs. of  8 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3
2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
##  $ X       : logi  NA NA NA NA NA NA ...
```

```r
# Fit the Multiple Linear Regression model
mlr_model <- lm(charges ~ age + sex + bmi + children + smoker + region, data
= df)
```

```r
# Summarize the model
summary(mlr_model)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     region, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```r
# Extract R-squared (R2) value from the summary
mlr_r2 <- summary(mlr_model)$r.squared
print(mlr_r2)
```

```
## [1] 0.750913
```

```r
# Predict on training data
mlr_predictions <- predict(mlr_model, newdata = df)

# Calculate Mean Squared Error (MSE) for the Multiple Linear Regression model
mlr_mse <- mean((df$charges - mlr_predictions)^2)
print(mlr_mse)
```

```
## [1] 36501893
```

```r
# Lasso Regression

# Prepare the predictor matrix and response vector
X <- model.matrix(charges ~ age + sex + bmi + children + smoker + region,
data = df)[, -1]
```

```r
y <- df$charges

# Check the dimensions of X and y
print(dim(X))
```

```
## [1] 1338    8
```

```r
print(length(y))
```

```
## [1] 1338
```

```r
# Fit the Lasso regression model using glmnet
lasso_model <- glmnet(X, y, alpha = 1)  # alpha = 1 specifies Lasso
regression

# Perform cross-validation to select lambda (regularization parameter)
cv_model <- cv.glmnet(X, y, alpha = 1)

# Print the cross-validation results
print(cv_model)
```

```
##
## Call:  cv.glmnet(x = X, y = y, alpha = 1)
##
## Measure: Mean-Squared Error
##
##     Lambda Index  Measure      SE Nonzero
## min   75.5    53 37107423 3000843       6
## 1se  931.1    26 39906041 3251775       3
```

```r
# Extract the best lambda value from cross-validation
best_lambda <- cv_model$lambda.min

# Refit the model with the best lambda
lasso_model_best <- glmnet(X, y, alpha = 1, lambda = best_lambda)

# Calculate R-squared (R2) value for the best model
lasso_r2 <- cor(y, predict(lasso_model_best, s = best_lambda, newx = X))^2
print(lasso_r2)
```

```
##             s1
## [1,] 0.7505644
```

```r
# Predict on training data with the best lambda
lasso_predictions <- predict(lasso_model_best, s = best_lambda, newx = X)

# Calculate Mean Squared Error (MSE) for the Lasso regression model
lasso_mse <- mean((y - lasso_predictions)^2)
print(lasso_mse)
```

```
## [1] 36566523
```

```r
# Ridge Regression

# Fit the Ridge regression model using glmnet
ridge_model <- glmnet(X, y, alpha = 0)  # alpha = 0 specifies Ridge
regression

# Perform cross-validation to select lambda (regularization parameter)
cv_model <- cv.glmnet(X, y, alpha = 0)

# Print the cross-validation results
print(cv_model)
```

```
##
## Call:  cv.glmnet(x = X, y = y, alpha = 0)
##
## Measure: Mean-Squared Error
##
##     Lambda Index  Measure      SE Nonzero
## min    953   100 37659056 2815583       8
## 1se   2416    90 40149882 3011855       8
```

```r
# Extract the best lambda value from cross-validation
best_lambda <- cv_model$lambda.min

# Refit the model with the best lambda
ridge_model_best <- glmnet(X, y, alpha = 0, lambda = best_lambda)

# Calculate R-squared (R2) value for the best model
ridge_predictions <- predict(ridge_model_best, s = best_lambda, newx = X)
ridge_r2 <- cor(y, ridge_predictions)^2
print(ridge_r2)
```

```
##             s1
## [1,] 0.7508171
```

```r
# Calculate Mean Squared Error (MSE) for the Ridge regression model
ridge_mse <- mean((y - ridge_predictions)^2)
print(ridge_mse)
```

```
## [1] 37104117
```