Assignment 7

Nova Bradford

2024-05-16

```
# Load necessary libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(randomForest)
```

```
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
# Read the data
data <- read.csv("/Users/novabradford/Downloads/Assignment7-main/raw-data/cohort.csv")
```

```
# Display summary of the dataset
summary(data)
```

```
##      smoke            female           age            cardiac
##  Min.   :0.0000   Min.   :0.000   Min.   :18.00   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:30.00   1st Qu.:0.000
##  Median :0.0000   Median :0.000   Median :41.00   Median :0.000
```

```
##  Mean   :0.1016   Mean   :0.487   Mean   :41.47   Mean   :0.038
##  3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:53.00   3rd Qu.:0.000
##  Max.   :1.0000   Max.   :1.000   Max.   :65.00   Max.   :1.000
##       cost
##  Min.   : 8478
##  1st Qu.: 9389
##  Median : 9664
##  Mean   : 9672
##  3rd Qu.: 9925
##  Max.   :11326
```

```r
# Create a table describing the variables
variable_summary <- data.frame(
  Variable = names(data),
  Class = sapply(data, class),
  MissingValues = sapply(data, function(x) sum(is.na(x))),
  UniqueValues = sapply(data, function(x) length(unique(x)))
)

# Print variable summary table
print(variable_summary)
```

```
##         Variable   Class MissingValues UniqueValues
## smoke      smoke integer             0            2
## female    female integer             0            2
## age          age integer             0           48
## cardiac  cardiac integer             0            2
## cost        cost integer             0         1597
```

```r
# Perform a linear regression analysis
model <- lm(cost ~ age, data = data) # Replace Y with the dependent variable name
summary(model)
```

```
##
## Call:
## lm(formula = cost ~ age, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -882.6 -224.6  -28.7  180.7 1449.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8932.3873    14.6785  608.53   <2e-16 ***
## age           17.8414     0.3365   53.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 322.1 on 4998 degrees of freedom
## Multiple R-squared:   0.36,  Adjusted R-squared:  0.3599
## F-statistic:  2812 on 1 and 4998 DF,  p-value: < 2.2e-16
```

```r
# Plot the regression results
plot(data$age, data$cost, main = "Scatter plot with regression line",
     xlab = "Age", ylab = "Cost", pch = 19)
abline(model, col = "red", lwd = 2)
```