HEALTHCARE COST ANALYSIS

Enock Bereka

2024-11-25

```
Load the dataset and the required libraries
library(tidyverse)

Healthcare_insurance <- read_csv("Healthcare insurance.csv")

View(Healthcare_insurance)
health <- Healthcare_insurance

Exploratory data analysis
head(health)
```

```
## # A tibble: 6 × 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    19 female  27.9        0 yes    southwest  16885.
## 2    18 male    33.8        1 no     southeast   1726.
## 3    28 male    33          3 no     southeast   4449.
## 4    33 male    22.7        0 no     northwest  21984.
## 5    32 male    28.9        0 no     northwest   3867.
## 6    31 female  25.7        0 no     southeast   3757.
```

```
tail(health)
```

```
## # A tibble: 6 × 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    52 female  44.7        3 no     southwest  11412.
## 2    50 male    31.0        3 no     northwest  10601.
## 3    18 female  31.9        0 no     northeast   2206.
## 4    18 female  36.8        0 no     southeast   1630.
## 5    21 female  25.8        0 no     southwest   2008.
## 6    61 female  29.1        0 yes    northwest  29141.
```

```
str(health)
```

```
## spc_tbl_ [1,338 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age     : num [1:1338] 19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr [1:1338] "female" "male" "male" "male" ...
##  $ bmi     : num [1:1338] 27.9 33.8 33 22.7 28.9 ...
##  $ children: num [1:1338] 0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr [1:1338] "yes" "no" "no" "no" ...
##  $ region  : chr [1:1338] "southwest" "southeast" "southeast" "northwest"
...
```

```
##  $ charges : num [1:1338] 16885 1726 4449 21984 3867 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   sex = col_character(),
##   ..   bmi = col_double(),
##   ..   children = col_double(),
##   ..   smoker = col_character(),
##   ..   region = col_character(),
##   ..   charges = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
glimpse(health)
```

```
## Rows: 1,338
## Columns: 7
## $ age      <dbl> 19, 18, 28, 33, 32, 31, 46, 37, 37, 60, 25, 62, 23, 56,
27, 1…
## $ sex      <chr> "female", "male", "male", "male", "male", "female",
"female",…
## $ bmi      <dbl> 27.900, 33.770, 33.000, 22.705, 28.880, 25.740, 33.440,
27.74…
## $ children <dbl> 0, 1, 3, 0, 0, 0, 1, 3, 2, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0,
0, 0…
## $ smoker   <chr> "yes", "no", "no", "no", "no", "no", "no", "no", "no",
"no", …
## $ region   <chr> "southwest", "southeast", "southeast", "northwest",
"northwes…
## $ charges  <dbl> 16884.924, 1725.552, 4449.462, 21984.471, 3866.855,
3756.622,…
```

```
anyNA(health)
```

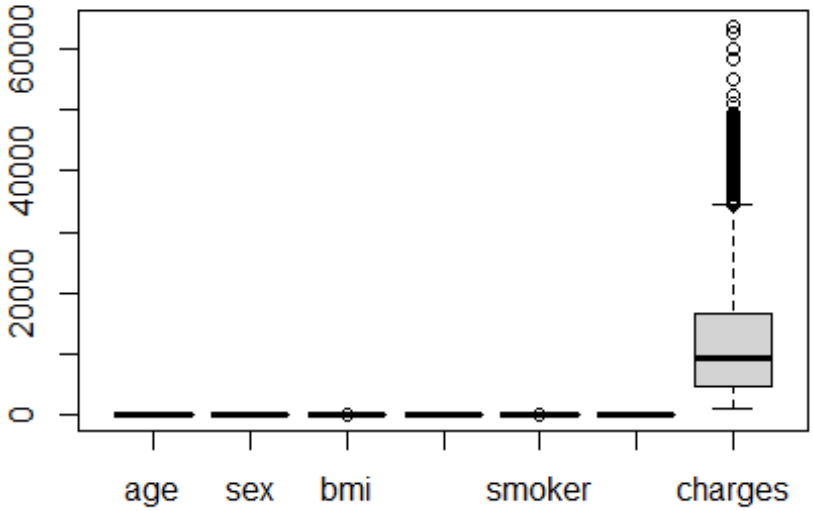
```
## [1] FALSE
```

setting up factors

```
health$smoker <- as.factor(health$smoker)
health$smoker <- as.numeric(health$smoker)
health$region <- as.factor(health$region)
health$region <- as.numeric(health$region)
health$sex <- as.factor(health$sex)
health$sex <- as.numeric(health$sex)
numeric_cols <- sapply(health, is.numeric)
```

Detecting and dealing with outliers

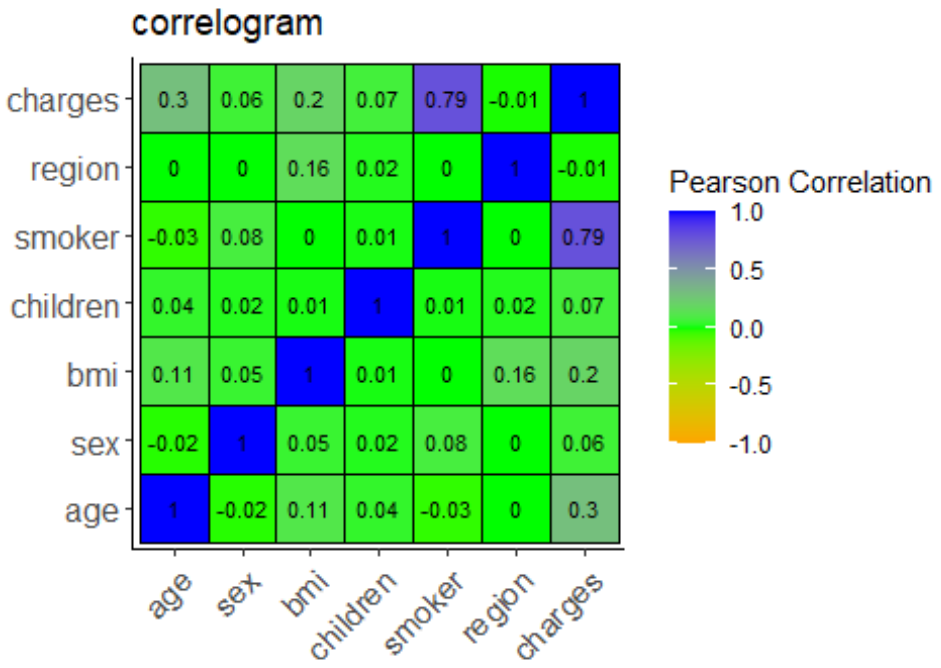
```
boxplot(health)
```




```
Checking the relationship between variables

corr <- cor(health)

Visualization of a correlation matrix
library(ggcorrplot)
ggcorrplot(corr,title = "correlogram",lab_col = "black",
           lab = TRUE, legend.title = "Pearson Correlation",
           lab_size = 3, ggtheme = theme_classic(),
           outline.color = "black",
           colors = c("orange", "green", "blue"))
```




```
Split the dataset into training and testing sets
library(caTools)
set.seed(123)
split <- sample.split(health$charges, SplitRatio = 0.7)
training <- subset(health, split == TRUE)
testing <- subset(health, split == FALSE)

Checking for multicollinearity
library(car)

model <- lm(charges~., data = training)
cor_matrix <- cor(health[, numeric_cols], use = "complete.obs")
print(cor_matrix)

##                  age          sex         bmi    children       smoker
## age       1.000000000 -0.020855872 0.109271882 0.04246900 -0.025018752
## sex      -0.020855872  1.000000000 0.046371151 0.01716298  0.076184817
## bmi       0.109271882  0.046371151 1.000000000 0.01275890  0.003750426
## children  0.042468999  0.017162978 0.012758901 1.00000000  0.007673120
## smoker   -0.025018752  0.076184817 0.003750426 0.00767312  1.000000000
## region    0.002127313  0.004588385 0.157565849 0.01656945 -0.002180682
## charges   0.299008193  0.057292062 0.198340969 0.06799823  0.787251430
##                region      charges
## age       0.002127313  0.299008193
## sex       0.004588385  0.057292062
## bmi       0.157565849  0.198340969
## children  0.016569446  0.067998227
```

```
## smoker   -0.002180682  0.787251430
## region    1.000000000 -0.006208235
## charges  -0.006208235  1.000000000
```

```
vif(model)
```

```
##      age      sex      bmi children   smoker   region
## 1.010673 1.021518 1.032853 1.003596 1.018062 1.021400
```

```
Implement our model
model <- lm(charges~., data = training)
summary(model)
```

```
##
## Call:
## lm(formula = charges ~ ., data = training)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11721  -3124  -1060   1861  29347
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35934.16    1434.15 -25.056   <2e-16 ***
## age            260.23      14.81  17.571   <2e-16 ***
## sex           -561.95     417.61  -1.346   0.1788
## bmi            365.53      34.09  10.723   <2e-16 ***
## children       535.76     171.12   3.131   0.0018 **
## smoker       24445.48     507.55  48.164   <2e-16 ***
## region        -464.80     188.26  -2.469   0.0137 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6320 on 929 degrees of freedom
## Multiple R-squared:  0.7536, Adjusted R-squared:  0.752
## F-statistic: 473.6 on 6 and 929 DF,  p-value: < 2.2e-16
```

(RSE): 6320 .This value measures the typical error in predicting charges. Lower values indicate better model fit.
Multiple R-squared: 0.7536 .This means approximately 75.36% of the variance in charges can be explained by the model. This is relatively high, suggesting a good fit.
Adjusted R-squared: 0.752. Adjusted R-squared accounts for the number of predictors, making it more reliable for multiple regression. Here, it's similar to R-squared, which confirms a strong model.
F-statistic: 473.6 on 6 and 929 DF, p-value < 2.2e-16. This indicates the model is statistically significant overall, meaning at least one predictor is significant in explaining charges

The model explains around 75% of the variance in charges.
Key predictors include age, BMI, number of children, smoking

status, and region. Among them, smoking has the largest positive effect on charges, suggesting it's a critical factor in predicting insurance costs. Other variables like sex show smaller and non-significant effects, meaning they don't substantially impact charges in this model.

```
Predicting the testing set
predictions <- predict(model, newdata = testing)
Compare predicted and actual values
results <- data.frame(Actual = testing$charges,
                      Predicted = predictions)
print(head(results))
```

```
##      Actual    Predicted
## 1  1725.552 3556.8059706
## 2 21984.471 3344.8194704
## 3  3866.855 5341.7079123
## 4  7281.506 8395.4003788
## 5  2721.321 3012.5728563
## 6  1837.237    0.3639274
```

```
head(testing$charges)
```

```
## [1]  1725.552 21984.471  3866.855  7281.506  2721.321  1837.237
```

```
head(predictions)
```

```
##            1            2            3            4            5
6
## 3556.8059706 3344.8194704 5341.7079123 8395.4003788 3012.5728563
0.3639274
```

```
Calculate Mean Squared Error (MSE)
mse <- mean((results$Actual - results$Predicted)^2)
cat("Mean Squared Error (MSE):", mse, "\n")
```

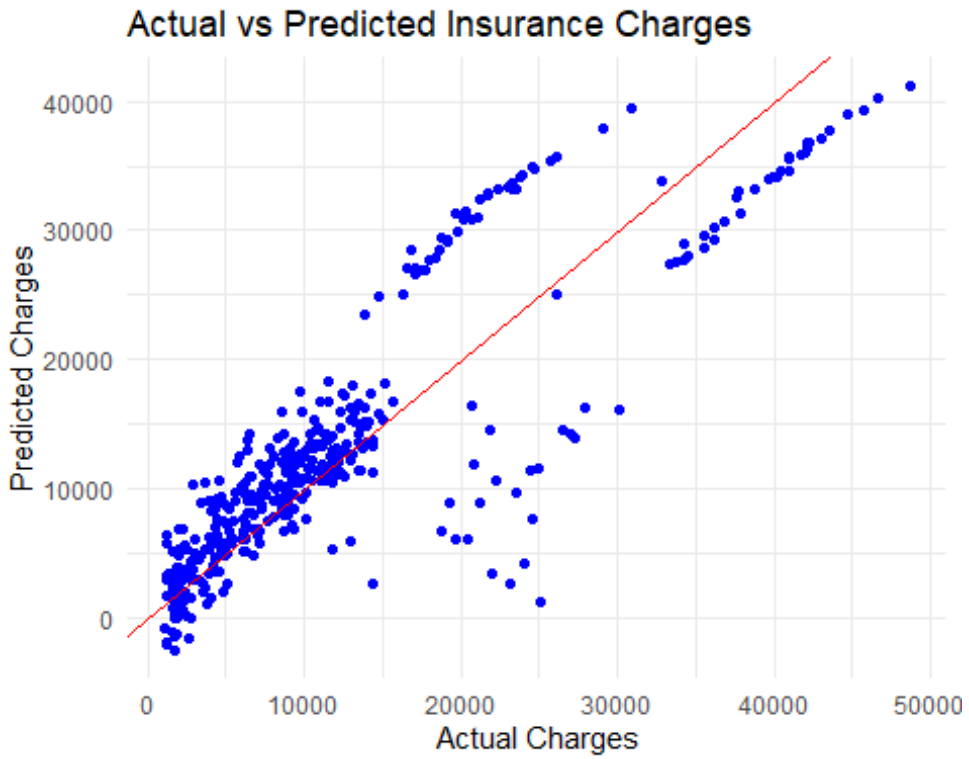
```
## Mean Squared Error (MSE): 29970718
```

```
Calculate R-squared for testing data
ss_total <- sum((testing$charges - mean(testing$charges))^2)
ss_residual <- sum((testing$charges - predictions)^2)
r_squared <- 1 - (ss_residual / ss_total)
cat("R-squared:", r_squared, "\n")
```

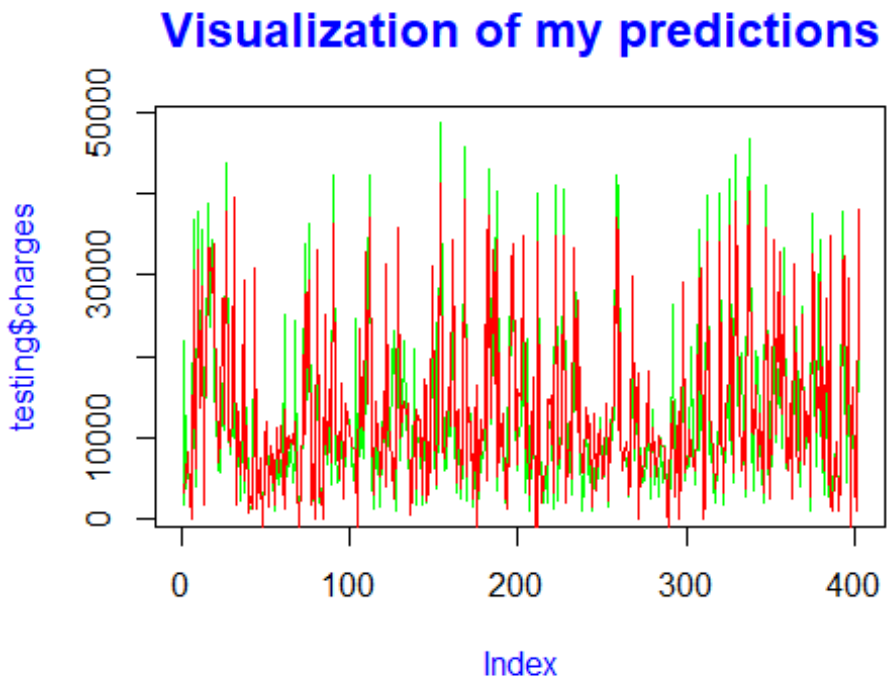
```
## R-squared: 0.7308464
```

```
Plot Actual vs. Predicted values
ggplot(results, aes(x = Actual, y = Predicted)) +
  geom_point(color = "blue") +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  labs(title = "Actual vs Predicted Insurance Charges",
       x = "Actual Charges",
```

```
       y = "Predicted Charges") +
  theme_minimal()
```




```
Another method
plot(testing$charges, type = "l", lty = 1.8,
     col = "green", col.main = "blue",
     main = "Visualization of my predictions",
     col.lab = "blue", cex.main = 1.5)
lines(predictions, type = "l", col = "red")
```


Visualization of my predictions
testing$charges
0
10000
30000
50000
0
100
200
300
400
Index