Medical Cost Analysis Project

Orator Murambiwa

2025-11-16

R Markdown

This project analyzes the Medical Cost Personal Dataset from Kaggle, which contains 1,338 patient records with variables such as age, sex, BMI, smoking status, number of children, region, and individual medical charges.

The goal of this project is to explore how demographic and lifestyle factors influence medical costs. The response variable for the regression model is charges, representing the total medical cost billed to an individual. Predictor variables include age, BMI, number of children, and smoking status

```
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)
```

```
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)
```

```
library(ggplot2)
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
```

Loading the Dataset

```
insurance <- read.csv("insurance.csv")
head(insurance)
```

```
##   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
```

Data Exploration

```r
summary(insurance)
```

```
##       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
sd(insurance$age)
```

```
## [1] 14.04996
```

```r
sd(insurance$bmi)
```

```
## [1] 6.098187
```

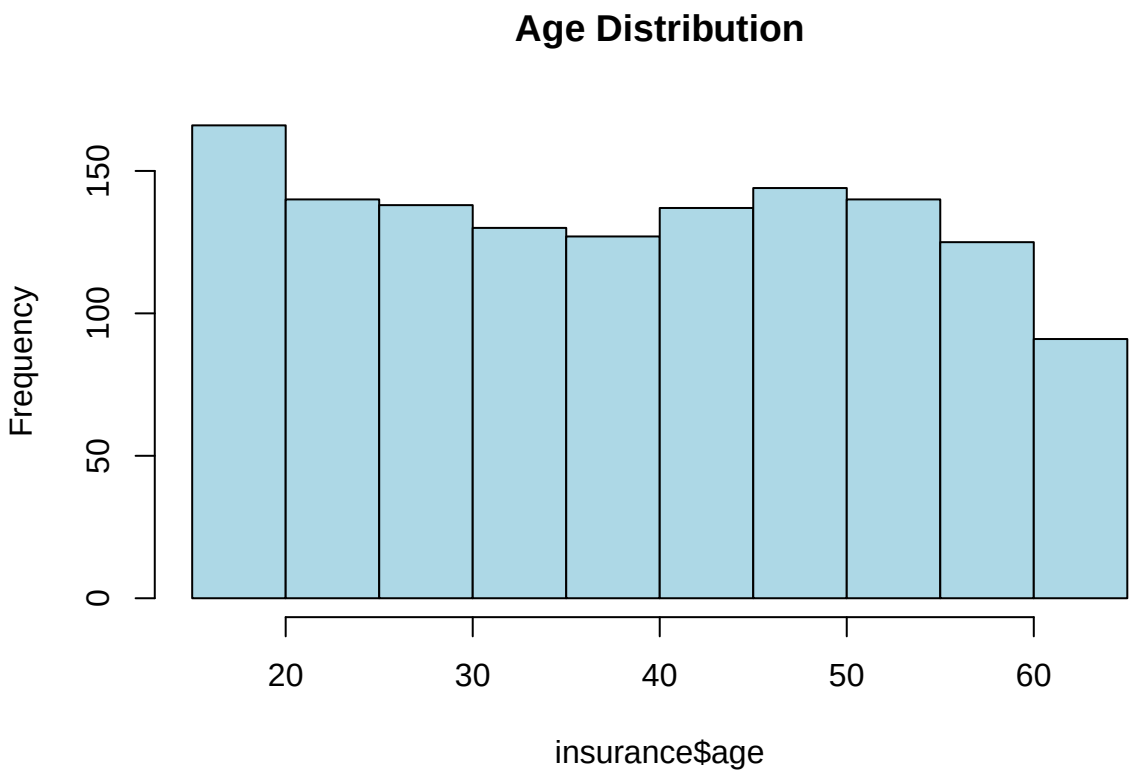
```r
sd(insurance$charges)
```

```
## [1] 12110.01
```

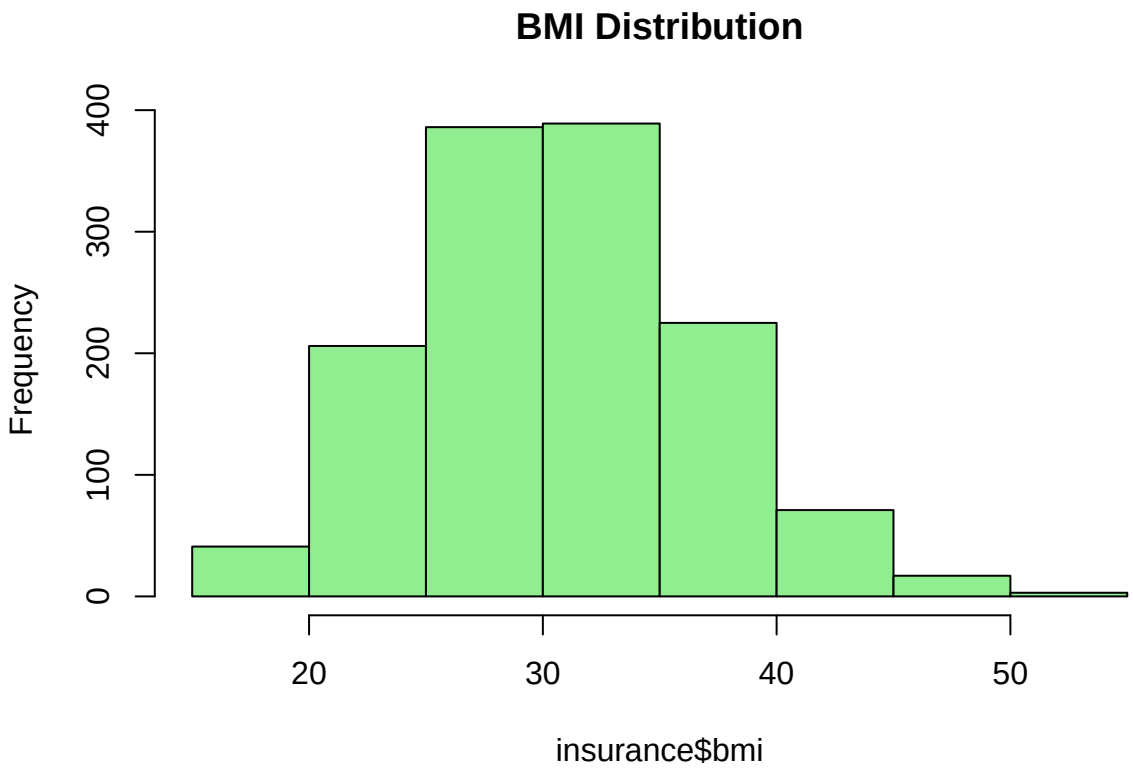
Creating Data Visualisations

Histograms

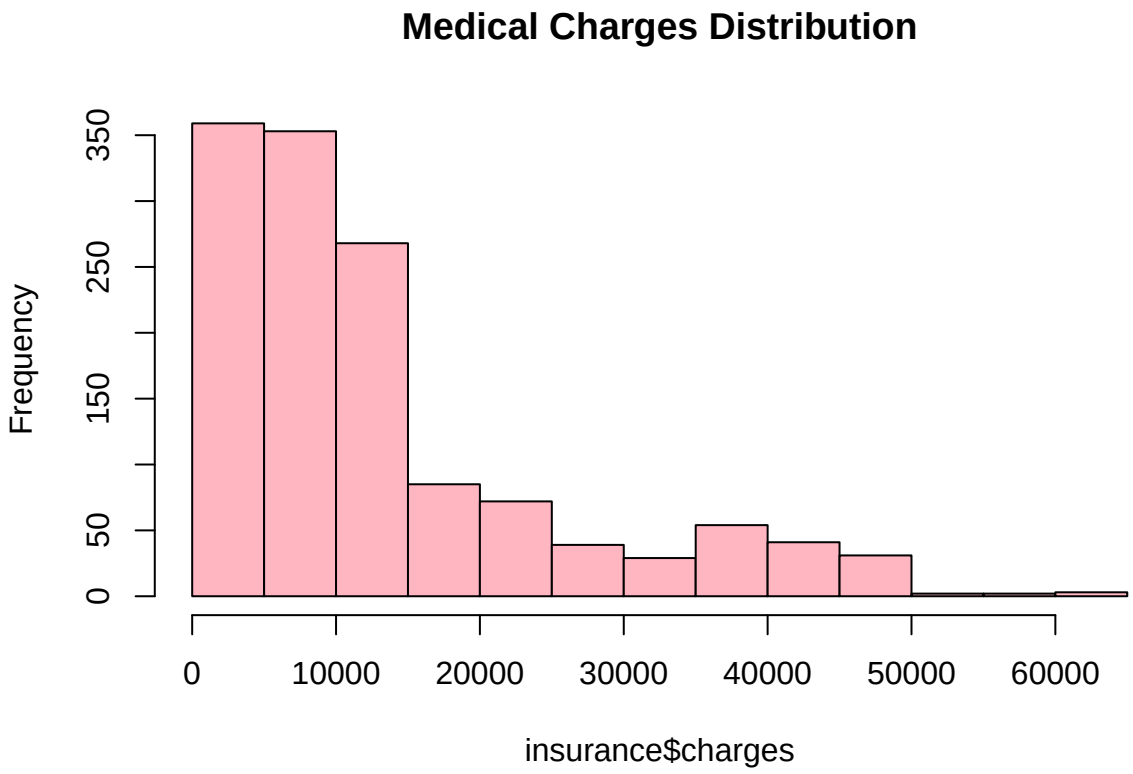
```r
hist(insurance$age, main="Age Distribution", col="lightblue")
```




```
hist(insurance$bmi, main="BMI Distribution", col="lightgreen")
```

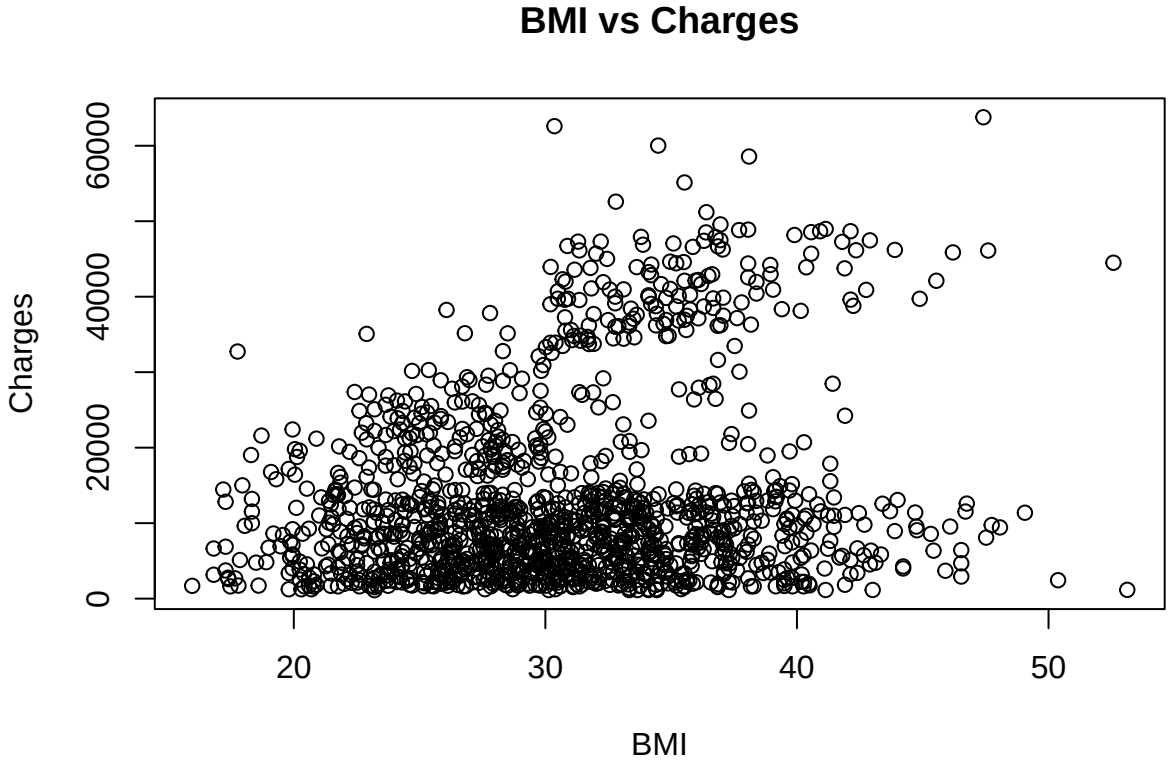



```
hist(insurance$charges, main="Medical Charges Distribution", col="lightpink")
```



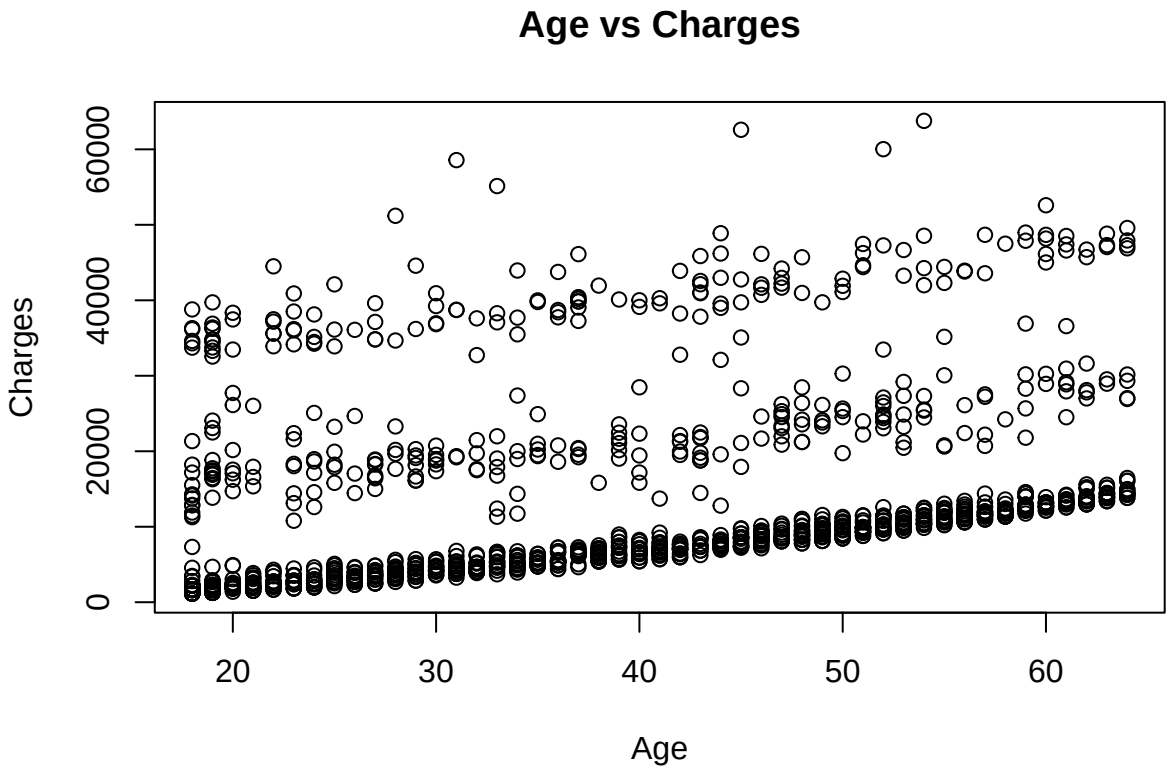
Scatter-plots

```
plot(insurance$bmi, insurance$charges,
     main="BMI vs Charges", xlab="BMI", ylab="Charges")
```



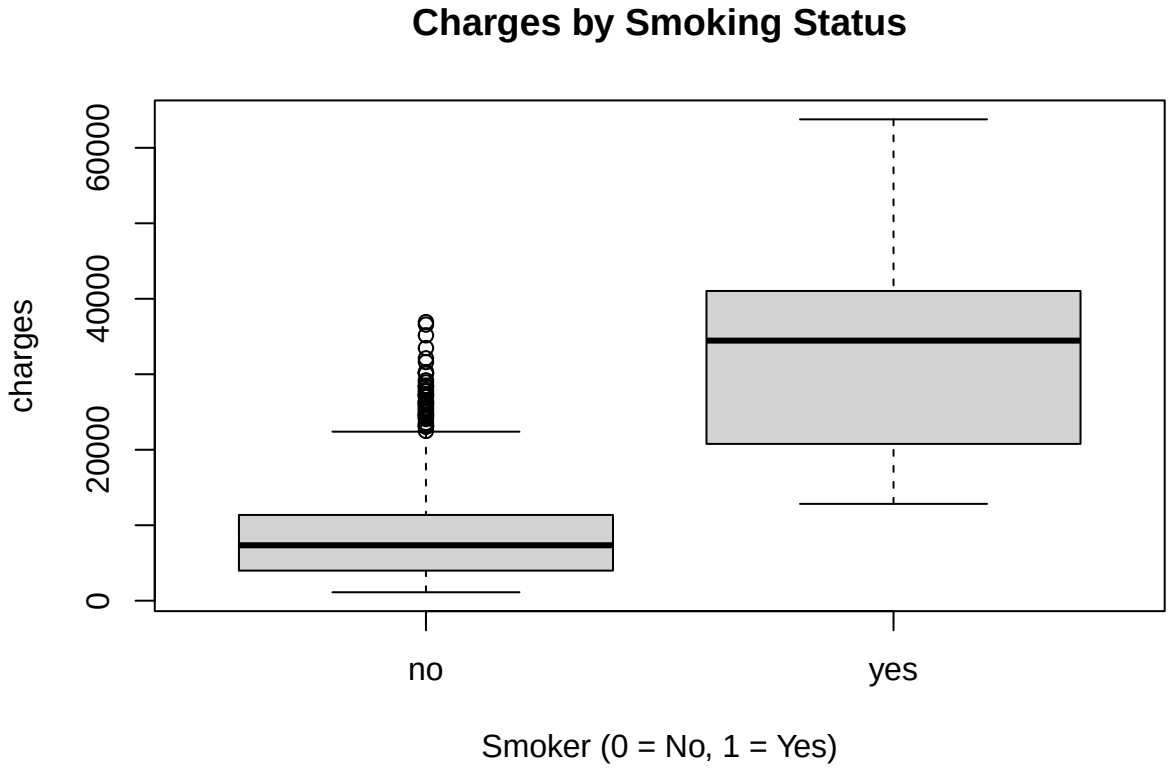

```
plot(insurance$age, insurance$charges,
```

```
     main="Age vs Charges", xlab="Age", ylab="Charges")
```



Boxplot(smoker status)

```
boxplot(charges ~ smoker, data=insurance,
        main="Charges by Smoking Status",
        xlab="Smoker (0 = No, 1 = Yes)")
```



Data Preprocessing

```
insurance$smoker <- ifelse(insurance$smoker == "yes", 1, 0)
insurance$sex <- ifelse(insurance$sex == "male", 1, 0)
```

Linear Regression Model

```
model <- lm(charges ~ age + bmi + children + smoker, data=insurance)
summary(model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11897.9  -2920.8   -986.6   1392.2  29509.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
## age            257.85      11.90  21.675  < 2e-16 ***
## bmi            321.85      27.38  11.756  < 2e-16 ***
## children       473.50     137.79   3.436 0.000608 ***
## smoker       23811.40     411.22  57.904  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6068 on 1333 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7489
## F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

Residual Plot

```
plot(model$residuals,
     main="Residual Plot",
     xlab="Index",
     ylab="Residuals")
abline(h=0, col="red")
```