Project_Health Insurance Analysis

Akshat Badhwar

2025-08-22

Importing the Dataset

```
setwd('C:/Users/aksha/Downloads')
insurance <- read.csv('insurance.csv')
```

Exploratory Data Analysis (EDA)

```
str(insurance) # Structure of dataset (Data types and Columns)
```

```
## '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 ...
```

```
summary(insurance) # Summary statistics of all variables
```

```
##       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
```

```
outliers <- boxplot.stats(insurance$charges)$out # Identifies outliers in
'Charges'
outliers
```

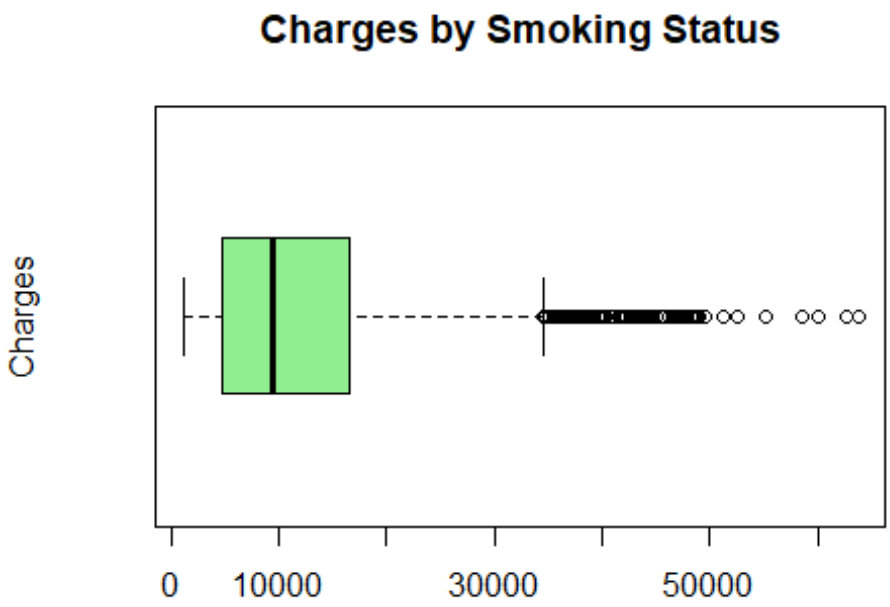
```
##   [1] 39611.76 36837.47 37701.88 38711.00 35585.58 51194.56 39774.28
48173.36
##   [9] 38709.18 37742.58 47496.49 37165.16 39836.52 43578.94 47291.06
47055.53
```

```
##  [17] 39556.49 40720.55 36950.26 36149.48 48824.45 43753.34 37133.90
34779.61
##  [25] 38511.63 35160.13 47305.31 44260.75 41097.16 43921.18 36219.41
46151.12
##  [33] 42856.84 48549.18 47896.79 42112.24 38746.36 42124.52 34838.87
35491.64
##  [41] 42760.50 47928.03 48517.56 41919.10 36085.22 38126.25 42303.69
46889.26
##  [49] 46599.11 39125.33 37079.37 35147.53 48885.14 36197.70 38245.59
48675.52
##  [57] 63770.43 45863.21 39983.43 45702.02 58571.07 43943.88 39241.44
42969.85
##  [65] 40182.25 34617.84 42983.46 42560.43 40003.33 45710.21 46200.99
46130.53
##  [73] 40103.89 34806.47 40273.65 44400.41 40932.43 40419.02 36189.10
44585.46
##  [81] 43254.42 36307.80 38792.69 55135.40 43813.87 39597.41 36021.01
45008.96
##  [89] 37270.15 42111.66 40974.16 46113.51 46255.11 44202.65 48673.56
35069.37
##  [97] 39047.29 47462.89 38998.55 41999.52 41034.22 36580.28 35595.59
42211.14
## [105] 44423.80 37484.45 39725.52 44501.40 39727.61 48970.25 39871.70
34672.15
## [113] 41676.08 44641.20 41949.24 36124.57 38282.75 46661.44 40904.20
36898.73
## [121] 52590.83 40941.29 39722.75 37465.34 36910.61 38415.47 41661.60
60021.40
## [129] 47269.85 49577.66 37607.53 47403.88 38344.57 34828.65 62592.87
46718.16
## [137] 37829.72 36397.58 43896.38
```

```
length(outliers) # Count of the number of outliers
```

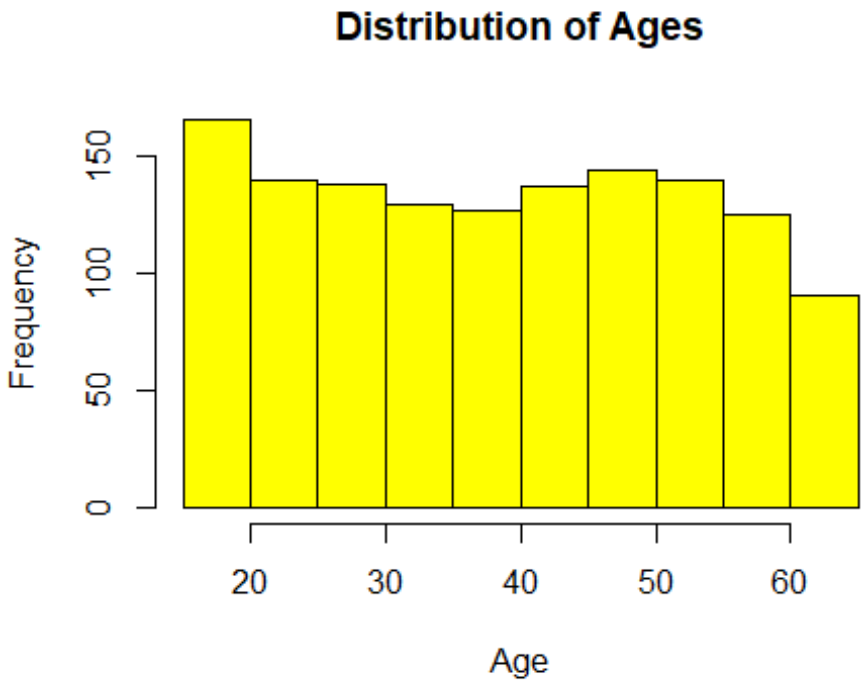
```
## [1] 139
```

```
boxplot(insurance$charges,
        main = "Charges by Smoking Status", ylab = "Charges",
        col = "lightgreen", horizontal = TRUE)    # Visualises outliers
```

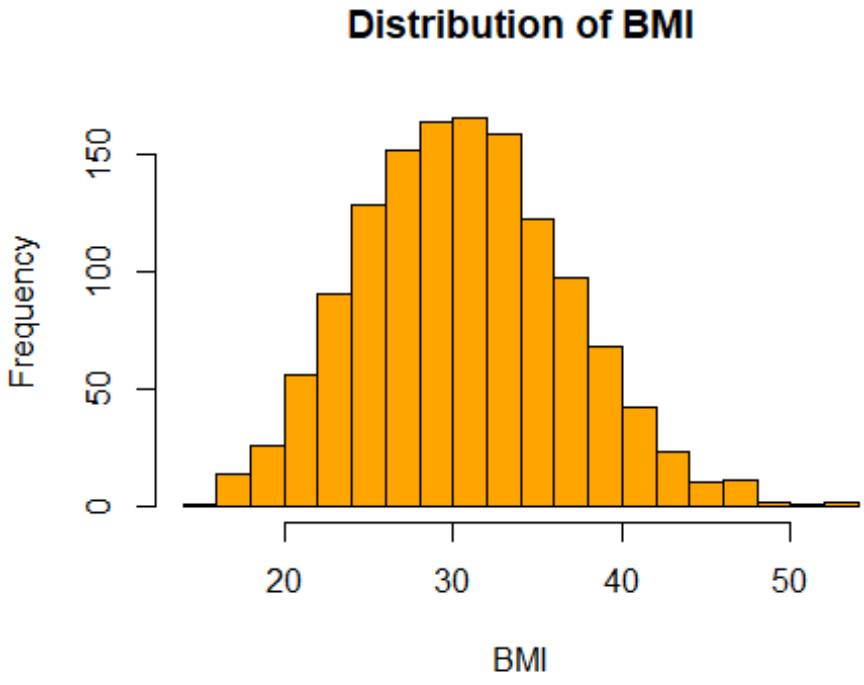


Descriptive Statistics and Visualisation

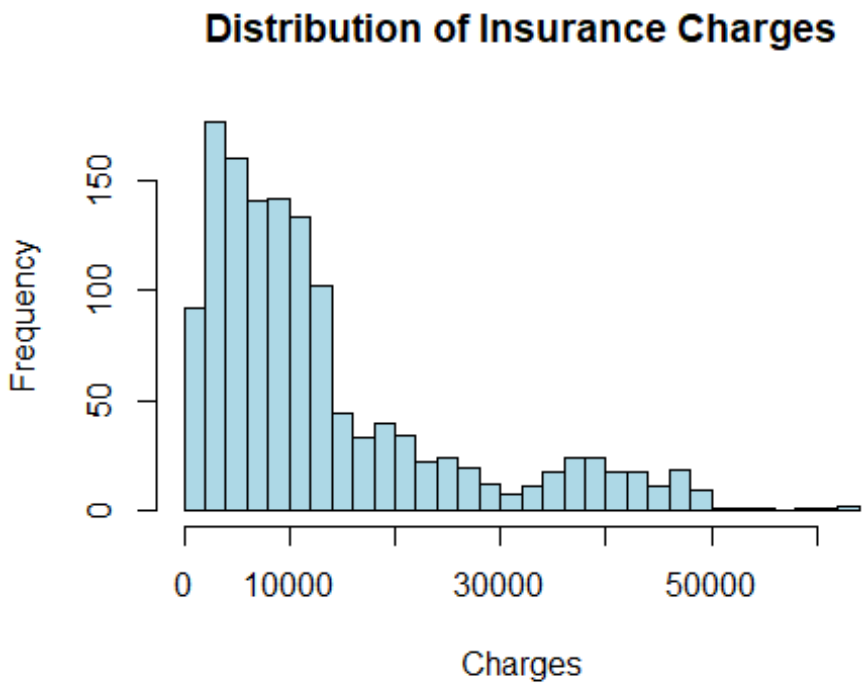
```
hist(insurance$age, main = 'Distribution of Ages', xlab = 'Age', col =
'yellow')
```



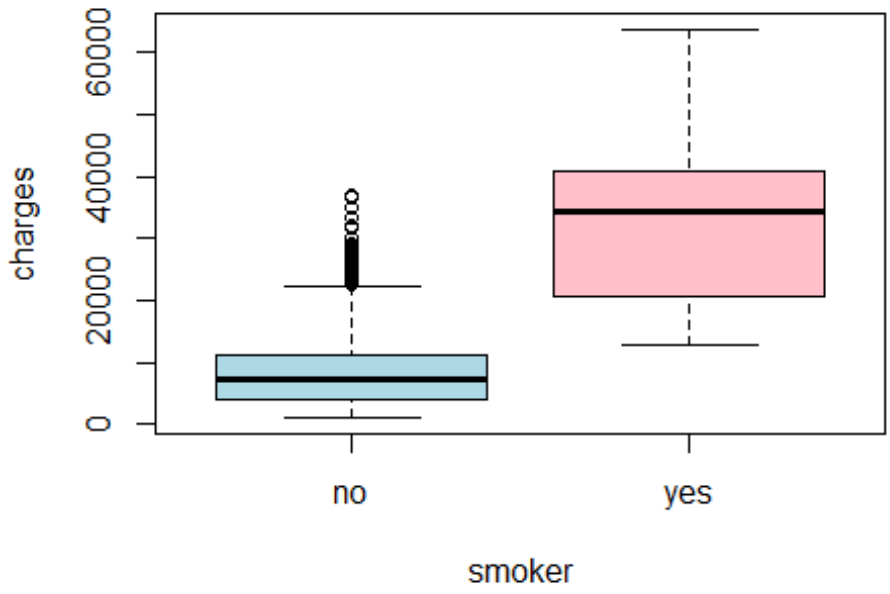

```r
hist(insurance$bmi, main = 'Distribution of BMI', xlab = 'BMI', col = 'orange', breaks = 20)
```



```r
hist(insurance$charges, main = 'Distribution of Insurance Charges', xlab =
'Charges', col = 'lightblue', breaks = 30)
```

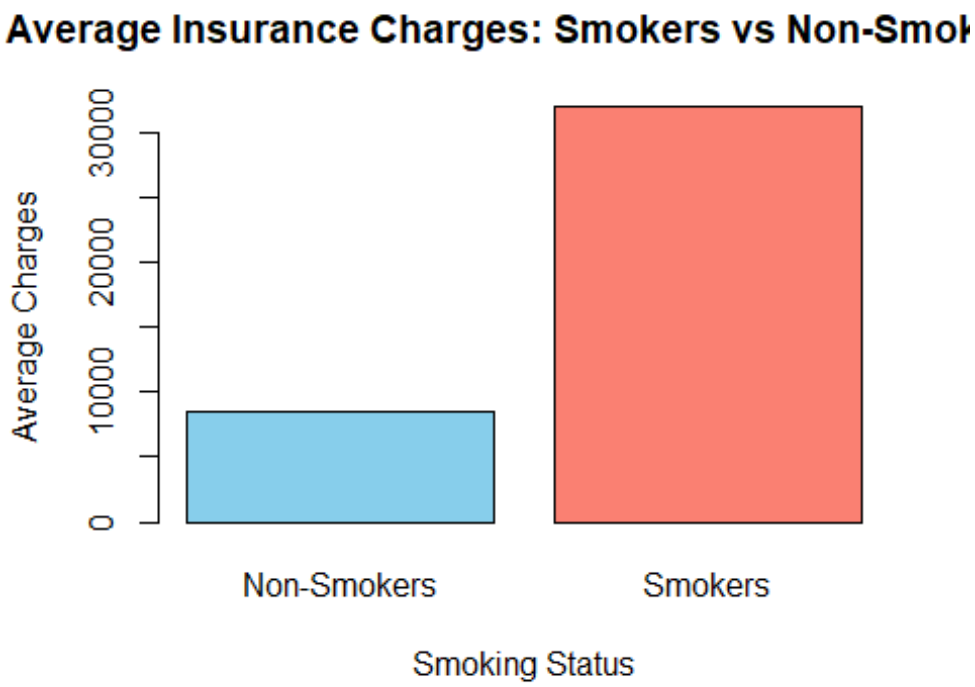



```r
boxplot(charges ~ smoker, data = insurance, col=c("lightblue","pink"))
```



Average Charges by Smoking Status

```r
insurance$smoker <- trimws(tolower(insurance$smoker)) # Cleans 'Smoker' variable

unique(insurance$smoker) # Verifies unique values

## [1] "yes" "no"

avg_charges <- tapply(insurance$charges, insurance$smoker, mean, na.rm = TRUE)

names(avg_charges) <- c("Non-Smokers", "Smokers")

barplot(avg_charges,
        main = "Average Insurance Charges: Smokers vs Non-Smokers",
        col = c("skyblue", "salmon"),
        ylab = "Average Charges",
        xlab = "Smoking Status")
```



Hypothesis Testing (Test 1: t-test)

H_0: Average charges for Smokers = Non-smokers H_1: Average charges for Smokers ≠ Non-smokers

```
smoker_charges <- insurance$charges[insurance$smoker == "yes"]

nonsmoker_charges <- insurance$charges[insurance$smoker == "no"]

t.test(charges ~ smoker, data = insurance)

##
##  Welch Two Sample t-test
##
## data:  charges by smoker
## t = -32.752, df = 311.85, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group no and
group yes is not equal to 0
## 95 percent confidence interval:
##  -25034.71 -22197.21
## sample estimates:
##  mean in group no mean in group yes
##          8434.268         32050.232
```

p-value < 0.05 (H0 is rejected) Average charges for Smokers is not equal to that of Non-Smokers

Hypothesis Testing (Test 2: ANOVA Test)

H_0: Mean Premiums are equal across regions H_1: Mean of at least one region is different

```
anova_result <- aov(charges ~ region, data=insurance)

summary(anova_result)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## region         3 1.301e+09 433586560    2.97 0.0309 *
## Residuals   1334 1.948e+11 146007093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(anova_result) # Identifies mean of which regions are significantly
different
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = charges ~ region, data = insurance)
##
## $region
##                             diff         lwr        upr     p adj
## northwest-northeast  -988.8091 -3428.93434 1451.31605 0.7245243
## southeast-northeast  1329.0269 -1044.94167 3702.99551 0.4745046
## southwest-northeast -1059.4471 -3499.57234 1380.67806 0.6792086
## southeast-northwest  2317.8361   -54.19944 4689.87157 0.0582938
## southwest-northwest   -70.6380 -2508.88256 2367.60656 0.9998516
## southwest-southeast -2388.4741 -4760.50957  -16.43855 0.0476896
```

p-value < 0.05 (H0 is rejected) Mean of at least one region is different

Hypothesis Testing (Test 3: Chi-Square Test)

H_0: Smoking and gender are independent (no relationship) H_1: Smoking and gender are dependent (there is a relationship)

```
smoke_gender <- table(insurance$smoker, insurance$sex)

chi_result <- chisq.test(smoke_gender)

chi_result
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  smoke_gender
## X-squared = 7.3929, df = 1, p-value = 0.006548
```

p-value < 0.05 (H0 is rejected) Smoking and gender are related

Correlation Analysis

```
round(cor(insurance[c("age","bmi","children","charges")]), 2) # Shows how strongly numeric variables are related
```

```
##           age  bmi children charges
## age      1.00 0.11     0.04    0.30
## bmi      0.11 1.00     0.01    0.20
## children 0.04 0.01     1.00    0.07
## charges  0.30 0.20     0.07    1.00
```

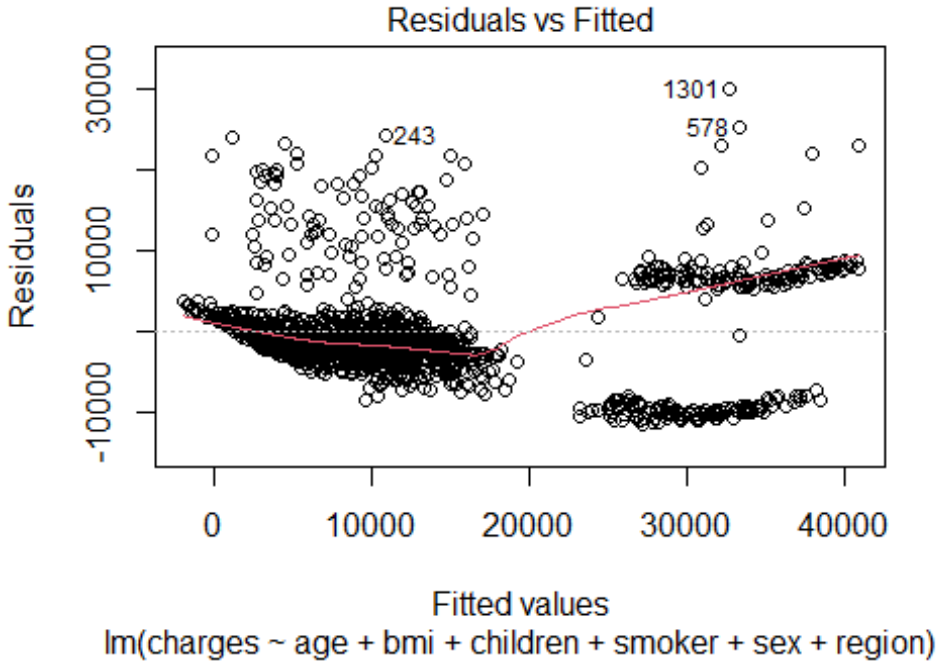
Regression Analysis and Model Diagnostics

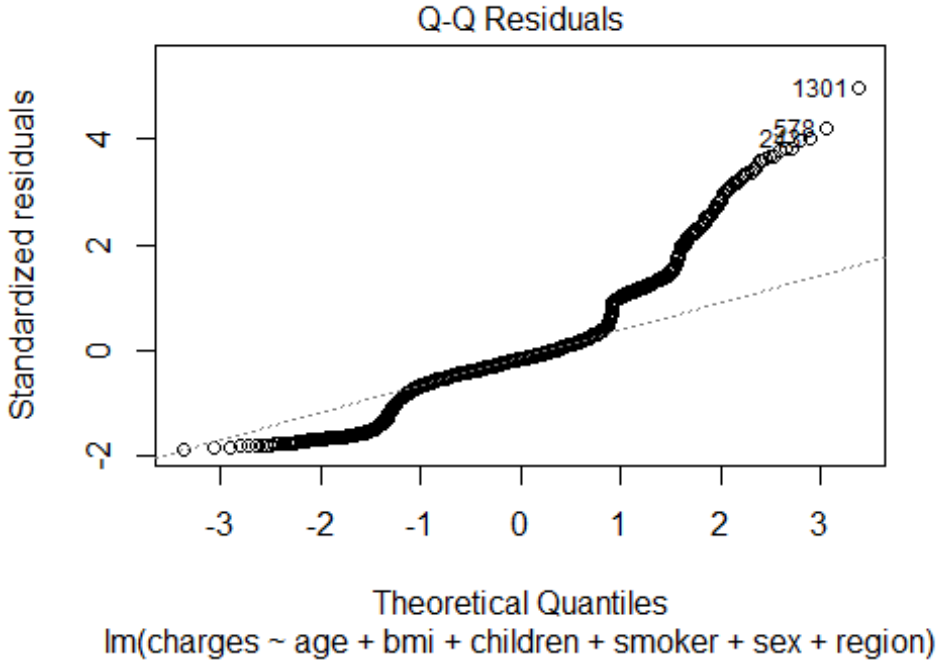
```
model <- lm(charges ~ age + bmi + children + smoker + sex + region, data=insurance)

summary(model) # Checks coefficients and significance
```

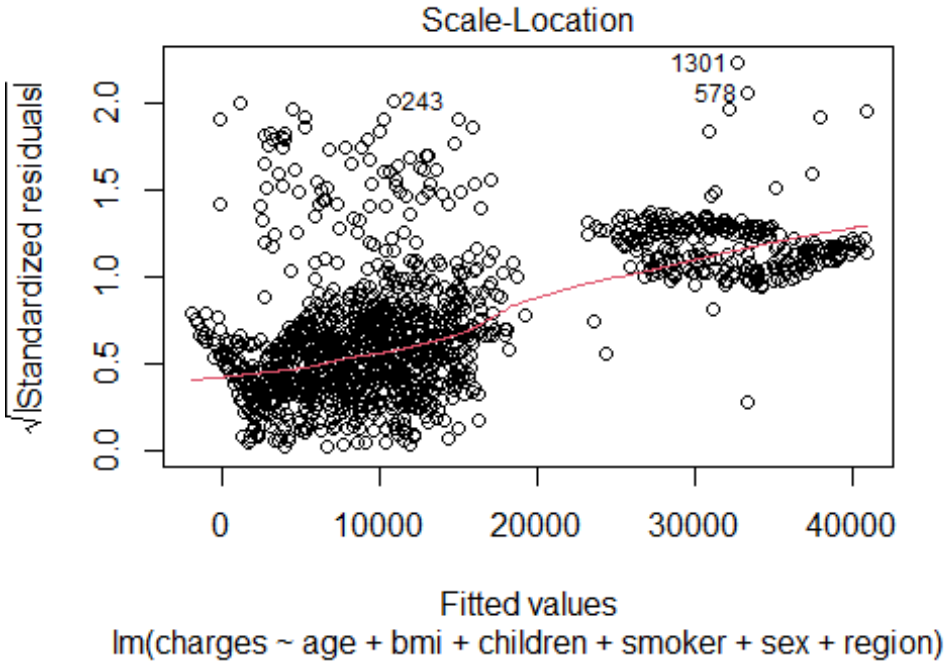
```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + sex +
##     region, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## 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 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## 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
```

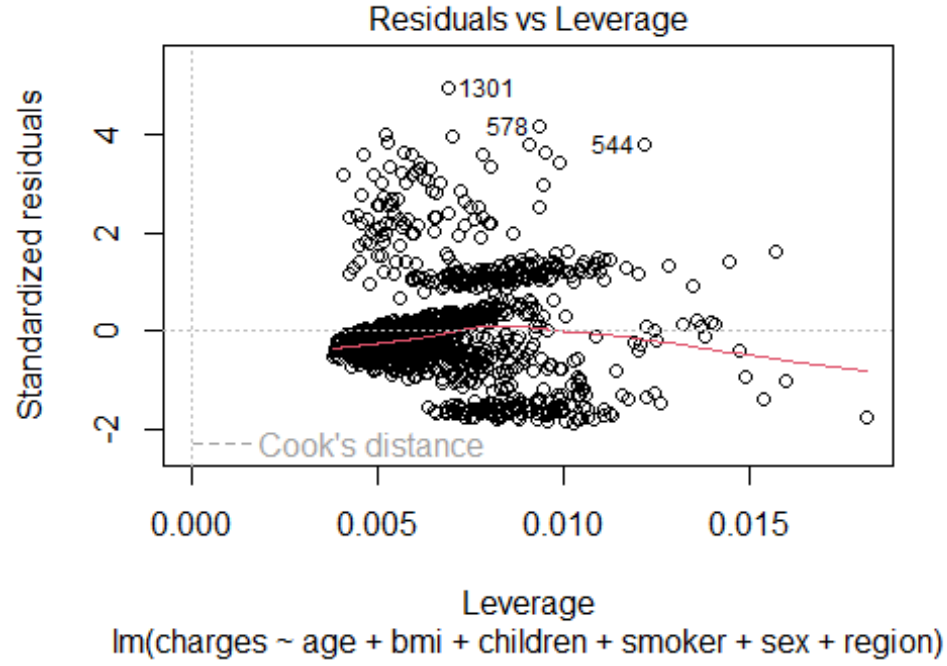
Diagnostic plots (Residuals, Normality, Leverage)

```
par(mfrow=c(2,2))

plot(model)
```


Residuals vs Fitted
Residuals
30000
10000
-10000
1301
578
243
0
10000
20000
30000
40000
Fitted values
lm(charges ~ age + bmi + children + smoker + sex + region)


Q-Q Residuals
Standardized residuals
4
2
0
-2
1301
578
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(charges ~ age + bmi + children + smoker + sex + region)






```
par(mfrow=c(1,1))
```