Price prediction

```
library(readxl)
data <- read_excel("insurance.xlsx")
head(data)
```

```
## # A tibble: 6 x 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.
```

#Create training and test set #Suffle the row and split the 80/20 train and test data

```
dt = sort(sample(nrow(data), nrow(data)*.8))
train<-data[dt,]
test<-data[-dt,]
```

```
dim(train)
```

```
## [1] 1070    7
```

```
dim(test)
```

```
## [1] 268   7
```

```
head(train)
```

```
## # A tibble: 6 x 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    28 male    33          3 no     southeast   4449.
## 3    33 male    22.7        0 no     northwest  21984.
## 4    32 male    28.9        0 no     northwest   3867.
## 5    46 female  33.4        1 no     southeast   8241.
## 6    60 female  25.8        0 no     northwest  28923.
```

```
head(test)
```

```
## # A tibble: 6 x 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    18 male    33.8        1 no     southeast   1726.
## 2    31 female  25.7        0 no     southeast   3757.
## 3    37 female  27.7        3 no     northwest   7282.
## 4    37 male    29.8        2 no     northeast   6406.
## 5    25 male    26.2        0 no     northeast   2721.
## 6    19 male    24.6        1 no     southwest   1837.
```

#train models

```
formula_0 <- as.formula("charges ~ age + sex + bmi + children + smoker + region")
model_0 <- lm(formula_0, data = train)
summary(model_0)
```

```
##
## Call:
## lm(formula = formula_0, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11244.3  -2932.8   -997.8   1423.7  30679.5
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11383.44    1120.93 -10.155  < 2e-16 ***
## age                249.60      13.50  18.493  < 2e-16 ***
## sexmale           -340.85     376.94  -0.904  0.36607
## bmi                340.04      32.19  10.563  < 2e-16 ***
## children           511.31     157.52   3.246  0.00121 **
## smokeryes        23651.68     468.57  50.477  < 2e-16 ***
## regionnorthwest   -564.70     538.78  -1.048  0.29483
## regionsoutheast  -1569.44     539.61  -2.908  0.00371 **
## regionsouthwest   -962.15     540.08  -1.781  0.07512 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6125 on 1061 degrees of freedom
## Multiple R-squared:  0.7397, Adjusted R-squared:  0.7378
## F-statistic: 376.9 on 8 and 1061 DF,  p-value: < 2.2e-16
```

train model without non-significant variable

```
formula_1 <- as.formula("charges ~ age + bmi + children + smoker + region")

model_1 <- lm(formula_1, data = train)
summary(model_1)
```

```
##
## Call:
## lm(formula = formula_1, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11043  -2921   -994   1354  30537
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11529.39    1109.15 -10.395  < 2e-16 ***
## age                249.99      13.49  18.534  < 2e-16 ***
## bmi                338.72      32.15  10.534  < 2e-16 ***
## children           507.08     157.44   3.221  0.00132 **
```

```
## smokeryes        23616.48     466.91  50.581  < 2e-16 ***
## regionnorthwest   -555.28     538.63  -1.031  0.30282
## regionsoutheast  -1564.39     539.53  -2.900  0.00381 **
## regionsouthwest   -947.64     539.80  -1.756  0.07945 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6125 on 1062 degrees of freedom
## Multiple R-squared:  0.7395, Adjusted R-squared:  0.7378
## F-statistic: 430.7 on 7 and 1062 DF,  p-value: < 2.2e-16
```