Medical Cost Analysis

Mark Arzola

```
knitr::opts_chunk$set(echo = TRUE)
```

```
install.packages("ggplot2", repos = "https://cran.r-project.org")
```

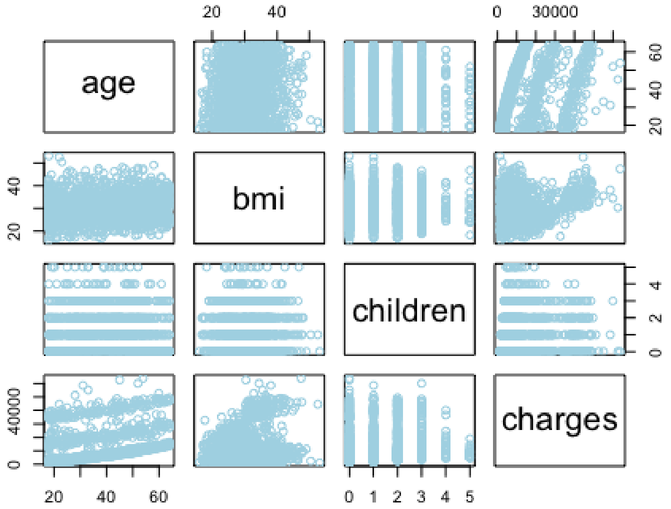
```
##
## The downloaded binary packages are in
##  /var/folders/3w/rdvgs5053xz_4sgwf7mp_p300000gn/T//Rtmp55VEqi/downloaded_packages
```

```
library(ggplot2)
```

```
setwd('/Users/markarzola/Desktop/portfolio projects/Insurance Analysis Project')
df <- read.csv('insurance.csv', header=TRUE)
head(df)
```

```
##   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
```

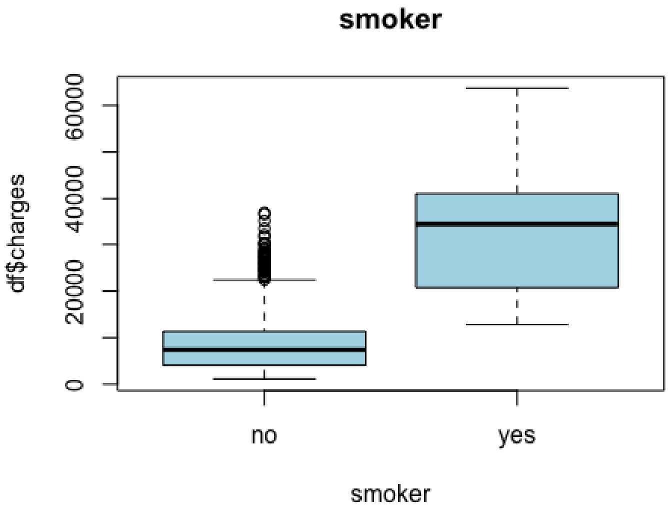
```
numericdata <- df[ , c(1,3,4,7)]
plot(numericdata, col = "lightblue")
```



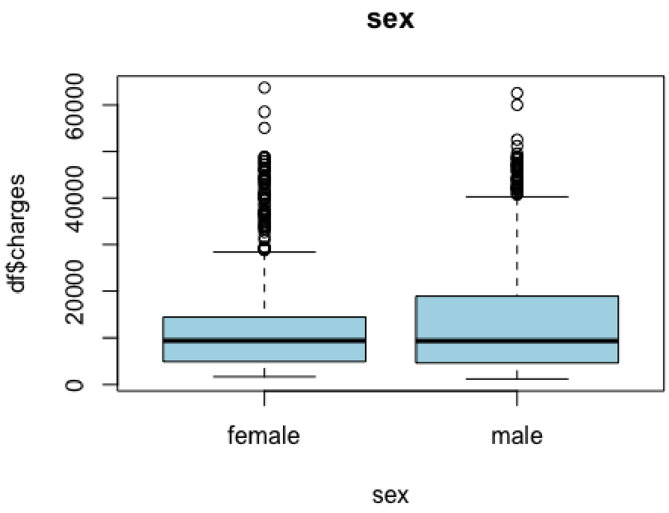
```r
round(cor(numericdata),2)
```

```
##           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
```

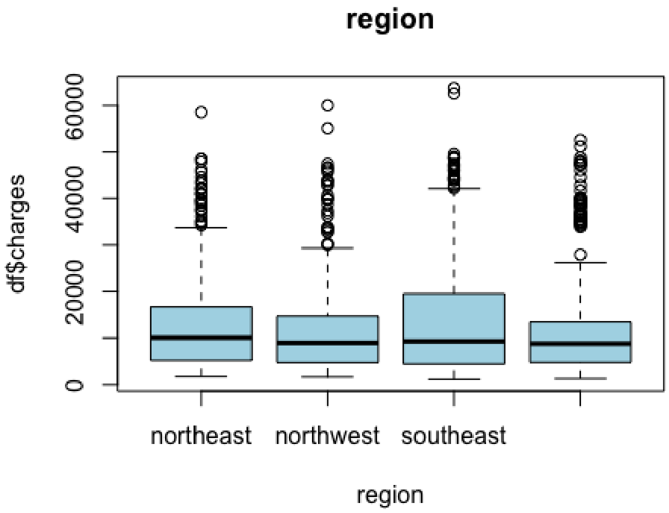
```r
smoker = as.factor(df$smoker)
sex = as.factor(df$sex)
region = as.factor(df$region)

boxplot(df$charges ~ smoker, main = 'smoker', col = "lightblue")
```



```r
boxplot(df$charges ~ sex, main = 'sex', col = "lightblue")
```



```
boxplot(df$charges ~ region, main = 'region', col = "lightblue")
```



```
model1 = lm(charges ~., data = df)
summary(model1)

##
## Call:
## lm(formula = charges ~ ., data = df)
##
## 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 ***
## sexmale            -131.3      332.9  -0.394 0.693348
## 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 ***
## 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
```