다변량통계학(1)_중간과제_201811530_통계학과_임도현

데이터 출처 : https://www.kaggle.com/datasets/mirichoi0218/insurance
(Book : Machine Learning with R by Brett Lantz)

선정한 데이터는 고객의 정보에 따른 의료비가 주어진 데이터 입니다.
이번 분석의 목표는 의료비와 고객들의 정보가 어떤 관계를 갖는지를 알아보는 것입니다.

```
# Check Data
data = read.csv("insurance.csv", header=TRUE)
X = data
head(X)
```

```
> head(X)
  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
```

dim(X)

```
> dim(X)
[1] 1338    7
```

summary(X)

```
> summary(X)
      age            sex                 bmi           children        smoker             region             charges     
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000   Length:1338        Length:1338        Min.   : 1122  
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000   Class :character   Class :character   1st Qu.: 4740  
 Median :39.00   Mode  :character   Median :30.40   Median :1.000   Mode  :character   Mode  :character   Median : 9382  
 Mean   :39.21                      Mean   :30.66   Mean   :1.095                                         Mean   :13270  
 3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000                                         3rd Qu.:16640  
 Max.   :64.00                      Max.   :53.13   Max.   :5.000                                         Max.   :63770  
```

변수설명
- age : 고객의 나이 (이산형 변수)
- sex : 고객의 성별 (명목형 변수)
- bmi : 고객의 bmi지수 (연속형 변수)
- children : 고객의 자녀 수 (이산형 변수)
- smoker : 고객의 흡연 여부 (명목형 변수)

- region : 고객의 거주 지역 (명목형 변수)
- charges : 고객의 의료비 (연속형 변수)

colSums(is.na(X)) # 결측치 존재 X

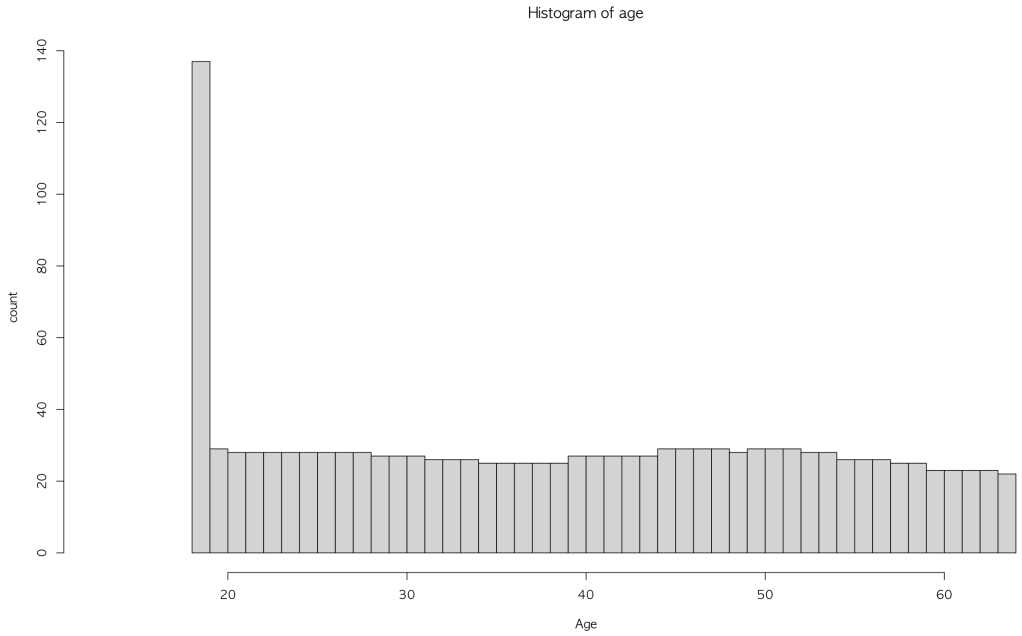
```
> colSums(is.na(X))
     age      sex      bmi children   smoker   region  charges
       0        0        0        0        0        0        0
```

데이터의 변수들을 구체적으로 확인해 봅시다.

Check $Age
summary(X$age)

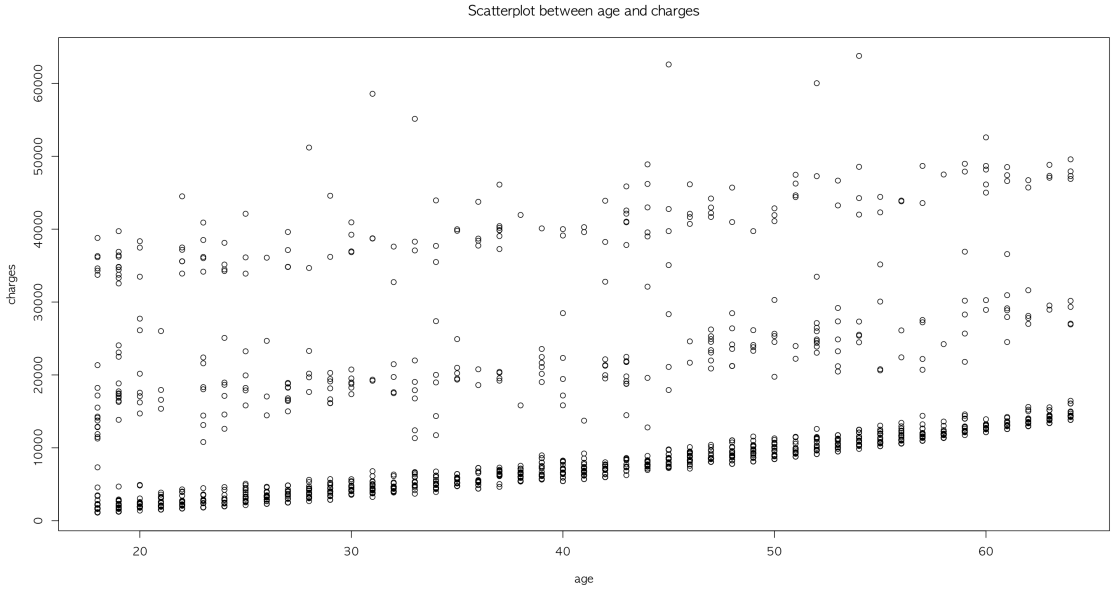
```
> summary(X$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   27.00   39.00   39.21   51.00   64.00
```

par(family="AppleGothic")
hist(X$age, main="Distribution of age", xlab="Age", ylab="count", xlim=c(min(X$age)-5, max(X$age)+3), breaks=length(unique(X$age)))



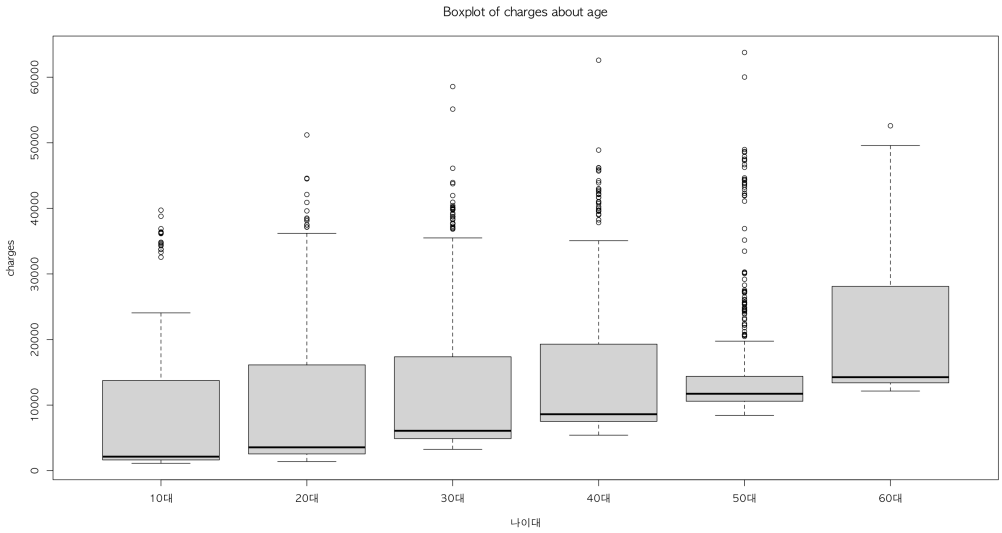
주어진 데이터에 10대 후반의 고객들이 다른 나이대의 고객들에 비해 상당히 많이 존재함을 알 수 있습니다.
고객들의 나이와 의료비의 관계를 산점도로 나타내어 봅시다.
plot(Xage, Xcharges, main="Scatterplot between age and charges", xlab="age",

ylab="charges")



전체적으로 나이가 증가할수록 의료비가 증가함을 알 수 있습니다. 직관적으로 와닿지 않으므로 나이를 나이대별로 grouping 후 boxplot을 살펴봅시다.

```
X2 = X[,c("age", "charges")]

for (i in 1:6){
  X2[(X2$age >= i*10) & (X2$age < (i+1)*10), "age"] = sprintf("%s0대", i)
}

boxplot(X2$charges~X2$age, xlab="나이대", ylab="charges", main="Boxplot of charges about age")
```



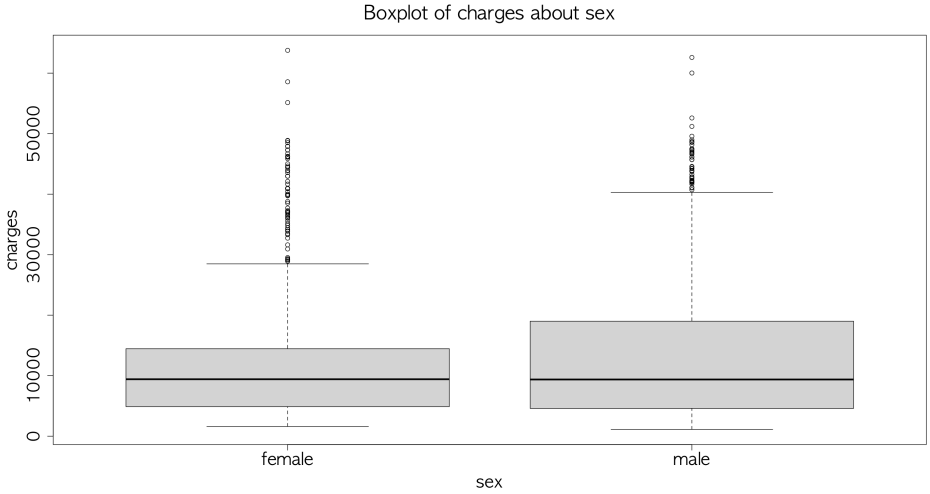
평균적으로 나이가 많을수록 의료비의 인상이 있습니다. 주목할점은 50대의 의료비의 분산이 비교적 다른 그룹에 비해 작으며 60대의 의료비는 이상치가 거의 존재하지 않음을 알 수 있습니다.

Check $sex
table(X$sex)

```
> table(X$sex)

female   male
   662    676
```

남녀의 수가 골고루 분포되어 있음을 알 수 있습니다.
성별과 의료비의 관계를 boxplot으로 확인해봅시다.

boxplot(X$charges~X$sex, xlab="sex", ylab="charges", main="Boxplot of charges about sex", cex.lab=1.8, cex.main=2, cex.axis=1.8)



의료비의 분산이 남성이 여성에 비해 크게 나타납니다. 또한 여성과 남성의 의료비는 평균적으로는 비슷하지만, 상위값이 남자들이 더 많이 분포되어 있는 것을 알 수 있습니다. 즉 남성들이 여성에 비해 의료비를 더 많이 내는 것으로 해석 가능합니다. 건강상태와 의료비는 매우 큰 관계가 있습니다. 우리의 데이터에서는 bmi지수와 흡연여부가 건강상태에 관련된 지표인데, 두 개의 변수의 남녀 비율을 확인해봅시다.

table(X[X$smoker=="yes", "sex"])

```
> table(X[X$smoker=="yes", "sex"])

female   male
   115    159
```

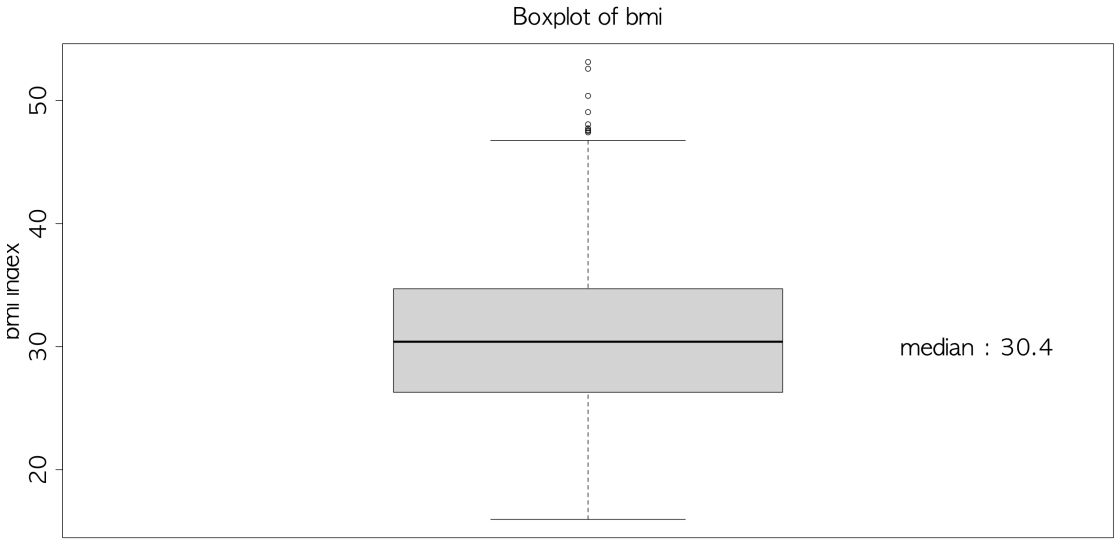
흡연자의 수가 여자보다 남자가 더 많이 존재합니다. 흡연자의 경우 비흡연자 보다 건강상태가 나쁘게 측정될 가능성이 높으므로, 남성의 의료비가 더 비싼 경향을 보인다고 추측할 수 있습니다.

table(X[X$bmi > mean(X$bmi), "sex"])

```
> table(X[X$bmi > mean(X$bmi), "sex"])

female   male
   306    340
```

비만으로 분류될 수 있는 비율이 남자가 더 많이 존재합니다. 비만이 합병증을 유발하므로 건강상태가 나쁘게 측정될 가능성이 높으므로, 남성의 의료비가 더 비싼 경향을 보인다고 추측할 수 있습니다.

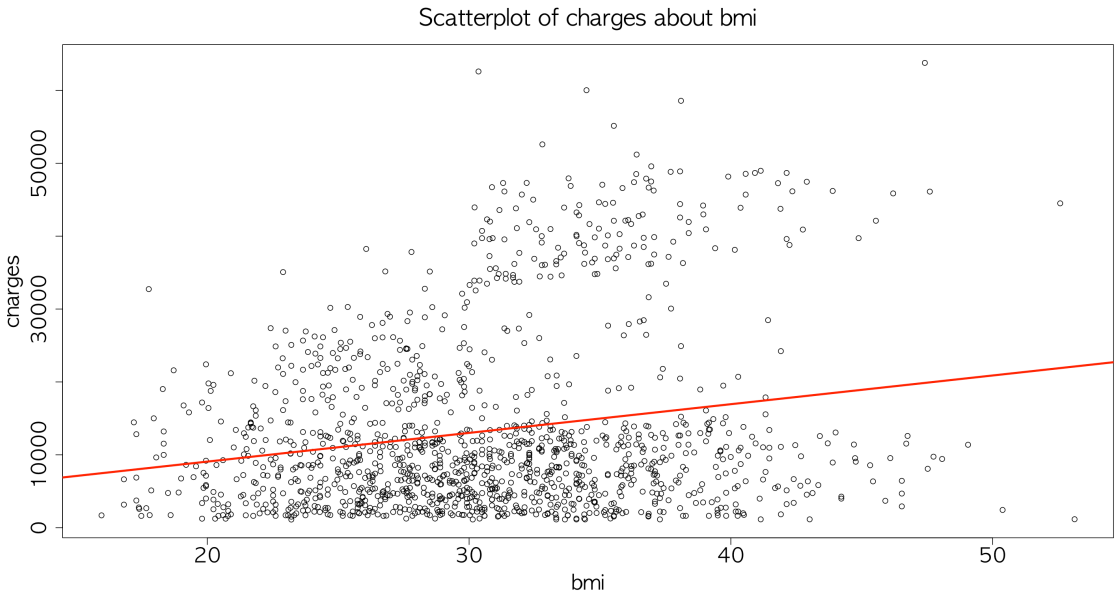
```
## Check $bmi
boxplot(X$bmi, main="Boxplot of bmi", ylab="bmi index", cex.main=2, cex.lab=2, cex.axis=2)
text(x=1.4, y=30, labels=paste("median :", round(median(X$bmi), 3)), cex=2)
```



중위수가 30.4로 고객들의 BMI 지수가 꽤 높아 보입니다. 또한 이상치로 간주되는 초고도비

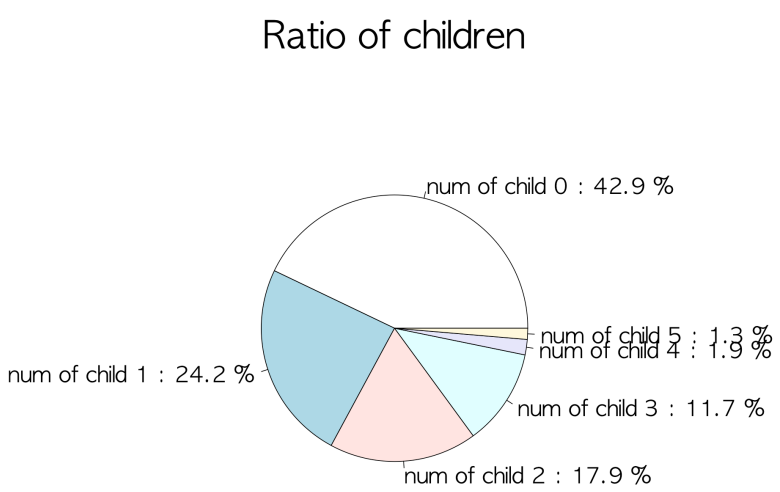
만의 고객들도 존재합니다.
고객들의 BMI 지수와 의료비의 관계를 살펴봅시다.

```
plot(X$charges~X$bmi, xlab="bmi", ylab="charges", main="Scatterplot of charges about bmi", cex.lab=1.8, cex.main=2, cex.axis=1.8)
abline(lm(X$charges ~ X$bmi), col="red", lwd=3)
```



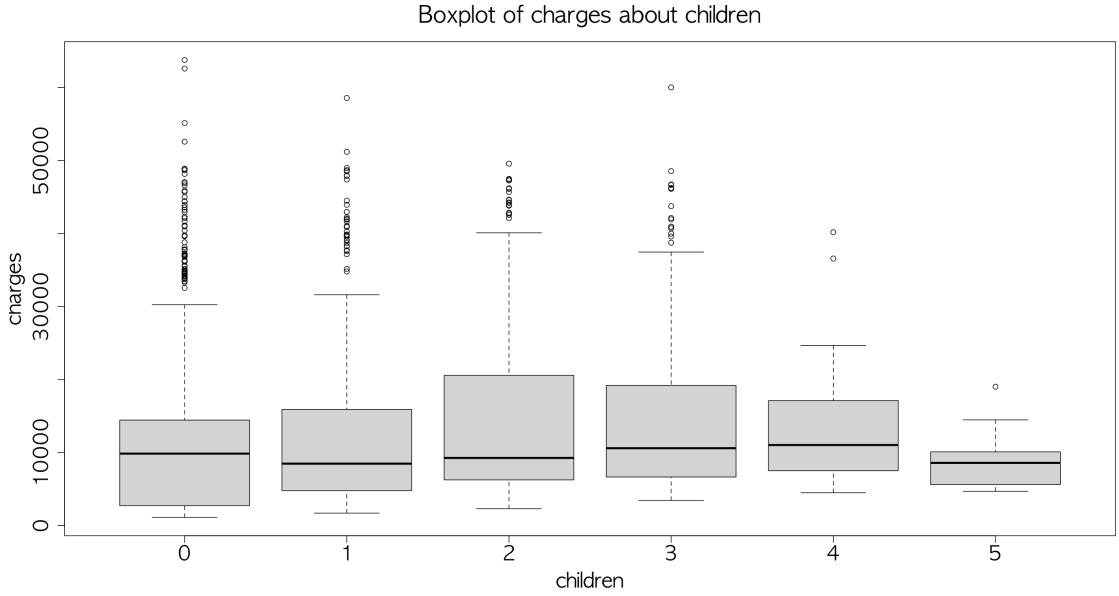
그래프에서 보이듯이 bmi 지수가 높아질수록 의료비가 높아짐을 알 수 있습니다. 비만이 각종 합병증을 유발하므로 의료비가 높게 측정되는 것으로 추측할 수 있습니다.

```
## Check $children
table(X$children)
tot.count.chil = sum(table(X$children))
obs.chil = paste("num of child", c("0","1","2","3","4","5"))

ratio.child = 0
for (i in 1:6){
  ratio.child[i] = 100*round(table(X$children)[i]/tot.count.chil, 3)
  }

pie(table(X$children), main="Ratio of children", labels=paste(obs.chil,":", ratio.child, "%"), cex=1.8, cex.main=3)
```



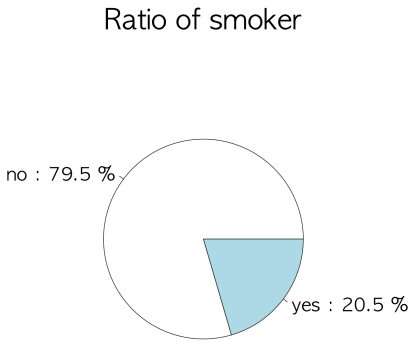
자녀를 가지지 않은 비율이 약 43%로 가장 많이 있음을 알 수 있습니다. 아마도 10대후반의 고객의 수가 다른 나이에 비해서 많기 때문이라고 유추할 수 있습니다.
자녀의 수와 의료비의 관계를 확인해봅시다.

boxplot(X$charges~X$children, xlab="children", ylab="charges", main="Boxplot of charges about children", cex.lab=1.8, cex.main=2, cex.axis=1.8)



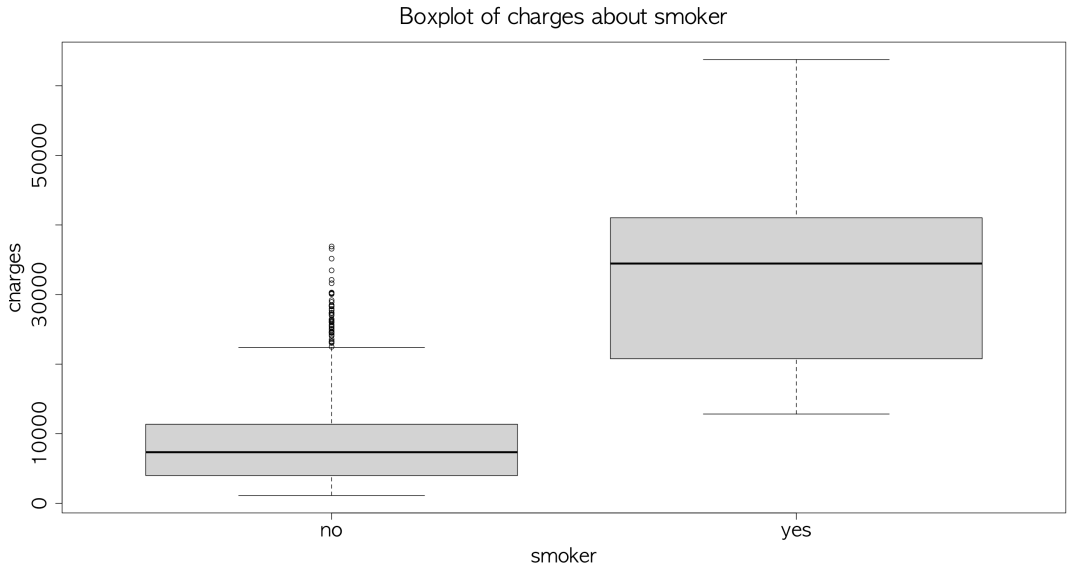
자녀의 수에 따른 의료비의 관계는 찾기 힘들어 보입니다.

```
## Check $smoker
table(X$smoker)
ratio.smoker = 100*round(table(X$smoker)/sum(table(X$smoker)), 3)

pie(table(X$smoker), main="Ratio of smoker", labels=paste(rownames(ratio.smoker), ":", ratio.smoker, "%"), cex=2, cex.main=3)
```



비흡연자의 비율이 80프로로 매우 높은걸 알 수 있습니다.
흡연여부와 의료비와의 관계를 확인해봅시다.

```
boxplot(X$charges~X$smoker, xlab="smoker", ylab="charges", main="Boxplot of charges about smoker", cex.lab=1.8, cex.main=2, cex.axis=1.8)
```



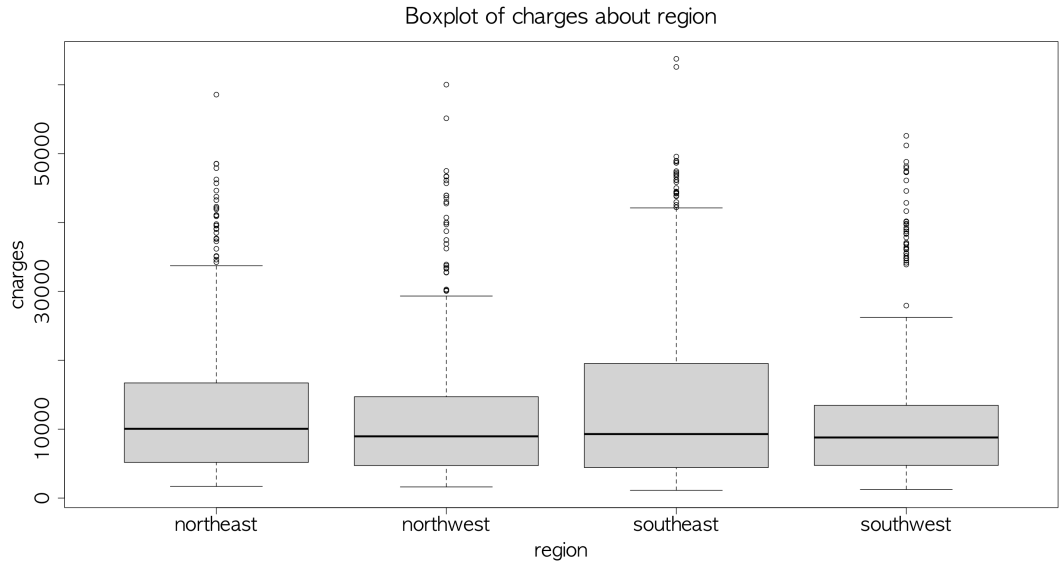
흡연자의 경우 의료비가 비흡연자에 비해 매우 높게 측정됨을 알 수 있습니다. 직관적인 결과입니다.

Check$region
table(X$region)

```
> table(X$region)

northeast northwest southeast southwest
      324       325       364       325
```

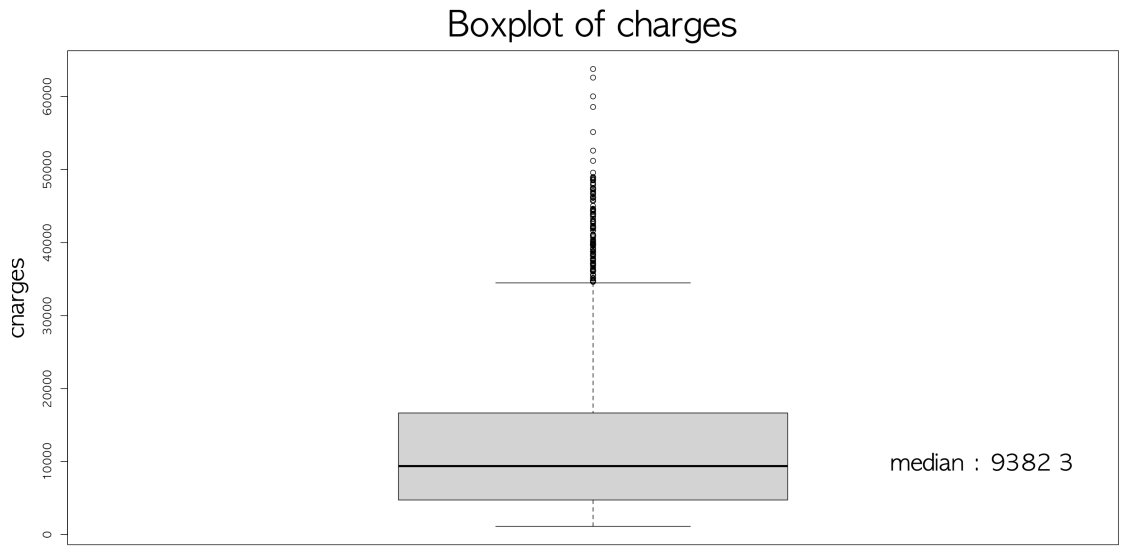
북동,북서, 남동, 남서 지방의 수가 골고루 분포되어 있음을 알 수 있습니다.
지역별 의료비의 관계를 확인해봅시다.

boxplot(X$charges~X$region, xlab="region", ylab="charges", main="Boxplot of charges about region", cex.lab=1.8, cex.main=2, cex.axis=1.8)



지역별로 의료비가 어떤 관계를 갖는지는 해석하기 조금 어려워 보입니다.

Check $charges
boxplot(X$charges, main="Boxplot of charges", ylab="charges", cex.main=3, cex.lab=2)
text(x=1.4, y=10000, paste("median :", labels=round(median(X$charges)), 3), cex=2)

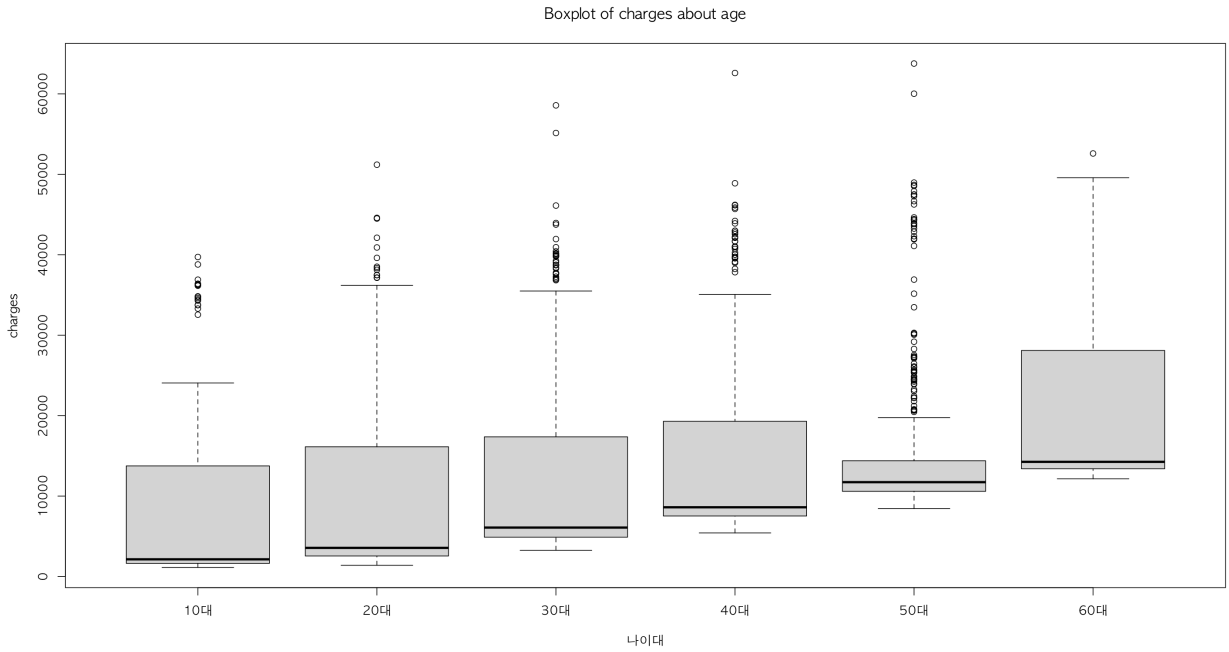


의료비의 중위수가 약 9382정도임을 알 수 있습니다. 하지만 이상치가 매우 많이 존재하며, 그래프의 y축 스케일이 매우 차이 나는 것을 보니 고객별로 의료비가 매우 상이함을 알 수 있습니다.

```
# Check the correlationships between variables and charges
cont.columns = c("age", "bmi", "children", "charges")
cor(X[,cont.columns])
```

```
> cor(X[,cont.columns])
                age       bmi    children     charges
age       1.0000000 0.1092719 0.04246900 0.29900819
bmi       0.1092719 1.0000000 0.01275890 0.19834097
children 0.0424690 0.0127589 1.00000000 0.06799823
charges   0.2990082 0.1983410 0.06799823 1.00000000
```

의료비에 가장 큰 영향을 미치는 변수는 나이임을 알 수 있습니다.



앞에서 확인하였던 나이대별 의료비 관계입니다. 나이대가 증가할수록 의료비도 증가하는 것을 직관적으로 확인할 수 있고 양의 상관관계 또한 확인할 수 있습니다.

```
# Check the indepedence of categorical variables
cat.columns = c("sex", "children", "smoker", "region")
summary(table(X[, cat.columns]))
```

```
> summary(table(X[, cat.columns]))
Number of cases in table: 1338
Number of factors: 4
Test for independence of all factors:
        Chisq = 93.74, df = 85, p-value = 0.2422
        Chi-squared approximation may be incorrect
```

chi-sqaure independence test 결과 p-value가 0.2422로서 범주형 자료들끼리의 독립성이 존재한다는 귀무가설을 기각하기 어려워 보입니다. 즉 범주형 자료들은 서로가 독립성을 만족한다고 결론 지을수 있습니다. 다음으로 관심있는 변수인 의료비에 따른 범주형 자료들의 독립성을 검정해봅시다.

```
table = xtabs(charges~sex+children+smoker+region, data=X)
summary(table)
```

```
> summary(table)
Call: xtabs(formula = charges ~ sex + children + smoker + region,
 data = X)
Number of cases in table: 17755825
Number of factors: 4
Test for independence of all factors:
        Chisq = 3079385, df = 85, p-value = 0
```

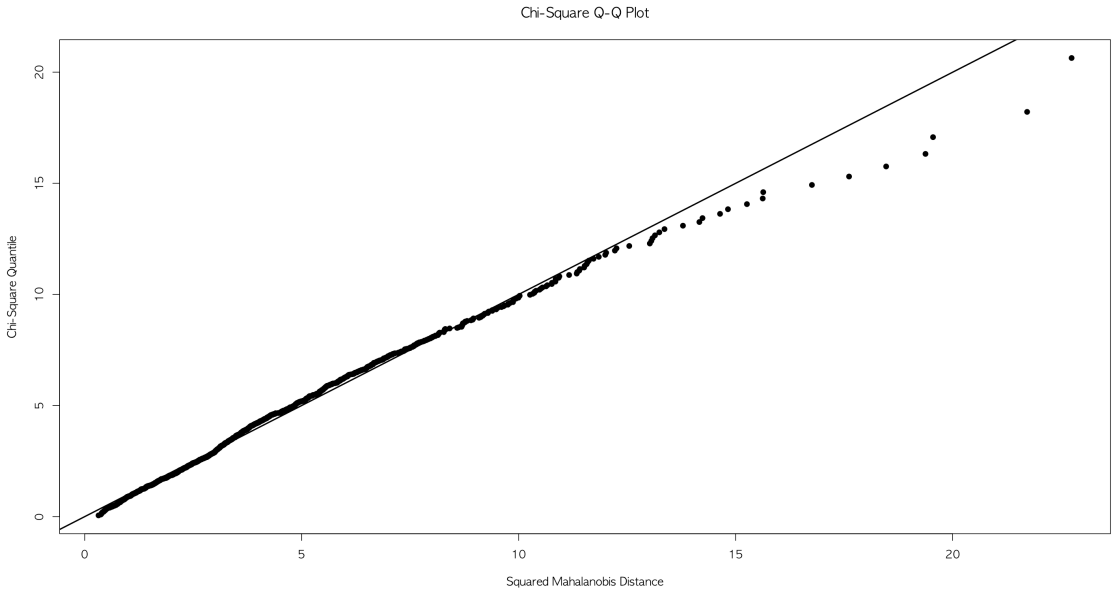
의료비에 따른 범주형 변수들의 독립성 검정에서 p-value가 0이므로 귀무가설인 독립성 존재를 기각할 수 있습니다. 즉 의료비에 대하여 범주형 변수들은 종속적인 관계를 갖음을 알 수 있습니다.

Check the normality

mardia-mvntest를 통해 연속형 변수들의 정규성을 확인해봅시다.

```
library(MVN)
mvn(X[,cont.columns], mvnTest="mardia",
multivariatePlot="qq")$multivariateNormality
```

```
> mvn(X[,cont.columns], mvnTest="mardia", multivariatePlot="qq")$multivariateNormality
             Test         Statistic               p value Result
1 Mardia Skewness  1007.10236610341 1.19383926312384e-200     NO
2 Mardia Kurtosis 0.622126891577761     0.533858433653274    YES
3             MVN              <NA>                  <NA>     NO
```



mardia normality test 결과 첨도(Kurtosis)부분에서는 정규성을 만족하지만 왜도(Skewness)부분에서 p-value가 매우 낮아 정규성을 만족하지않아 결론적으로 다변량정규성을 만족하지 못함을 알 수 있습니다.
Q-Q plot상으로 Mahalanobis Distance의 큰 값들이 정규성을 갖지 못하는 것에 영향을 주는것처럼 보입니다.