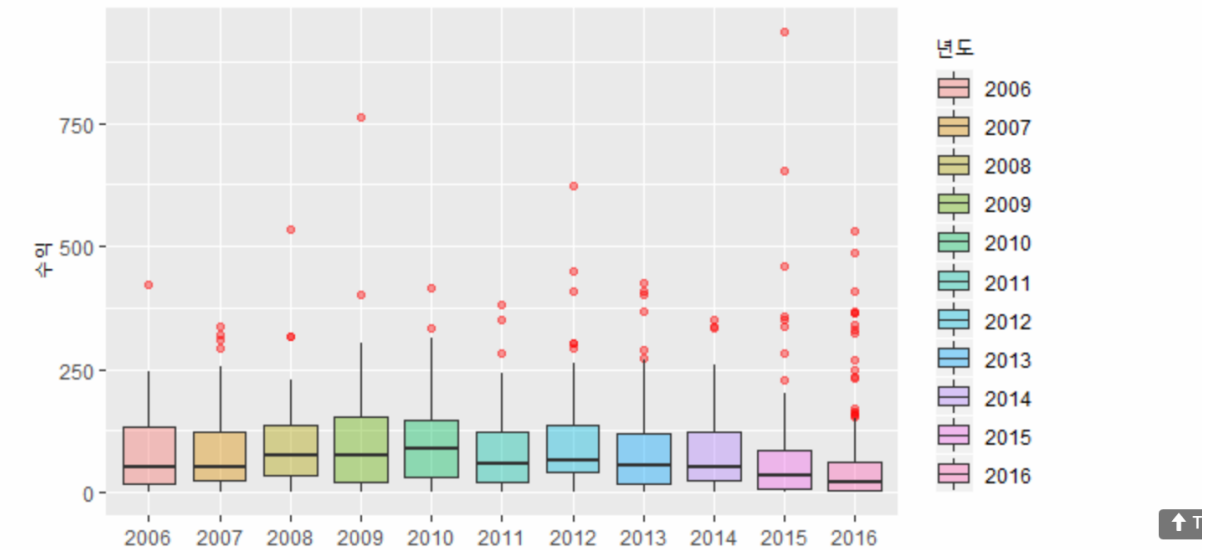
Ch3. 이상치(극단값, Outlier) 뽑아내기

이상치(Outlier)는 '패턴에서 벗어난 값'으로 정의를 내릴 수 있습니다. 또는 '중심에서 좀 많이 떨어져 있는 값' 이라고 할 수 있습니다. 이상치는 다음과 같은 특성을 지니고 있습니다.

- 평균에 막대한 영향을 미칩니다.
- [1,2,3,4,5] 의 평균은 3이지만, [1,2,3,4,100]의 평균은 22입니다.
- 하지만 중위수는 3으로 같습니다.
- 따라서 종종 평균으로 값을 나타내기보다는 중위수로 요약값을 나타내는 것이 더 현실을 반영하는 경우가 많습니다.

- 이상치 여부는 다음과 같은 식으로 계산을 합니다.

```
ggplot(IMDB,aes(x=as.factor(Year),y=Revenue..Millions.))+
  geom_boxplot(aes(fill=as.factor(Year)),outlier.colour = 'red',alpha=I(0.4))+
  xlab("년도") + ylab("수익") + labs(fill = "년도")
```



박스플롯을 그린 이유는 이상치 탐색은 바로 박스플롯으로 하기 때문입니다. 박스플롯은 다음과 같은 원리로 그려집니다.

- 박스플롯은 분위수를 기준으로 그려집니다.
- 상자 안에 그려져 있는 직선은 중위수(Median)을 나타냅니다.
- 상자의 밑변은 1분위수를 나타애며, 윗변은 3분위수를 나타냅니다.
- 상자를 중심으로 위 아래에, 직선이 있는 것을 볼 수 있습니다.
    - 이 직선은 "울타리"라고 부릅니다.

상자로부터 아래 직선의 계산식 : Q1 - 1.5 * (Q3 - Q1)

상자로부터 위 직선의 계산식 : Q3 + 1.5 *(Q3 - Q1)

- 이 울타리를 벗어난 값들을 Outlier라고 부릅니다.

- Outlier는 기본적으로 통계추정에 있어서 방해가 되고는 합니다.
- 통계분석은 전부 "귀납법"인데, 이상치같은 특수 케이스가 규칙을 만드는데 방해가 되기 때문입니다.

L`에시'로 부터 규칙을 생성한 증명하는 방법.

- Outlier의 처리방법
    - 제거를 하는 방법이 쓰이기는 하지만, 개인적으로 좋아하는 방식은 아닙니다. 어찌됐든 데이터를 버리는거니깐요.
    - 데이터 변형을 통해 Outlier문제를 줄여줍니다.
        - 통계추정에세는 정규분포를 맞추어 주는 것이 매우 중요합니다. 보통 Outlier로 인해 한 쪽으로 치우친 분포는 log 변환을 통해 정규성을 맞추어주고는 합니다. (이 부분은 후에 다루도록 하겠습니다.)

```
# Outlier인 데이터 제거하기

# 1분위수 계산
Q1 = quantile(IMDB$Revenue..Millions.,probs = c(0.25),na.rm = TRUE)
# 3분위수 계산
Q3 = quantile(IMDB$Revenue..Millions.,probs = c(0.75),na.rm = TRUE)

LC = Q1 - 1.5 * (Q3 - Q1) # 아래 울타리
UC = Q3 + 1.5 * (Q3 - Q1) # 위 울타리

IMDB2 = subset(IMDB,
               Revenue..Millions. >  LC & Revenue..Millions. < UC)
```