Boxplots

2016314726 정영준

Boxplot(상자 그림)은 기술통계학의 수치적 자료를 표현하기 위한 그래프 중 하나입니다. Boxplot을 그리는데 필요한 정보는 five-number summary입니다. Five-number summary란 5가지 요약 수치를 뜻하며 최솟값, 제1 사분위값, 중앙값, 제3 사분위값, 최댓값을 Five-number summary라고 합니다. 이 값들을 가지고 Boxplot을 그리는 법에 대해 설명하겠습니다. 먼저 제1 사분위값과 제3 사분위값을 양 끝으로 하는 Box를 그립니다. 그리고 Box내부에 중앙값의 위치를 표현합니다. 다음으로 outlier 판별을 위한 IQR을 계산하겠습니다. IQR은 제 3사분위수(Q3)에서 제 1사분위수(Q1)을 뺀 값을 의미하며 이는 앞서 그린 Box의 길이와 동일합니다. 그리고 IQR을 활용해서 이상치를 판별할 수 있습니다. Q1 – 1.5 \* IQR 보다 작거나 Q3 + 1.5 \* IQR보다 큰 값은 이상치로 판별이 가능합니다. 이상치로 판별된 값은 보통 o표로 Boxplot에 따로 표기를 합니다. 그리고 마지막으로 Q1 – 1.5 \* IQR과 같거나 크고 Q3 + 1.5 \* IQR보다 같거나 작은 값 중 최솟값과 최댓값을 실선으로 표시를 하고 이를 앞서 만든 Box와 이어서 그리면 Boxplot이 완성됩니다. Q1 – 1.5 \* IQR과 같거나 크고 Q3 + 1.5 \* IQR보다 같거나 작은 값 중 최솟값과 최댓값은 이상치를 제외한 값 중 최솟값과 최댓값이라고 해석이 가능합니다. 다음으로 Boxplot을 해석하는 과정에 대해 설명하겠습니다. 먼저 최솟값 최댓값을 표시한 꼬리부분이 box에서 길게 나와있으면 이는 데이터의 분포에서 꼬리부분이 길다고 할 수 있습니다. 그리고 중위수를 표시한 실선의 box내 위치에 따라 어느 곳에 데이터가 집중 분포해 있는지 알 수 있습니다. 만약 Box의 길이 즉 IQR값이 작다면 이는 적은 범위에 데이터가 전체의 50%가 존재한다는 의미합니다. Box의 길이가 길면 이는 전체적으로 데이터가 고르게 퍼져있다는 것을 뜻합니다. Boxplot을 사용하여 이상치를 검출하고 이에 대한 해석을 따로 하거나 머신러닝 분석을 위해 이상치를 제거할 수도 있습니다. 마지막으로 Boxplot의 장점과 단점에 대해 알아보겠습니다. Boxplot의 장점은 먼저 매우 많은 데이터를 한눈에 알아볼 수 있도록 하는 것 입니다. 줄기 잎 그림등의 다른 그래프에서는 많은 데이터를 처리하기가 힘듭니다. 하지만 Boxplot은 five-number summary를 통해 전체 데이터의 분포를 손쉽게 설명이 가능합니다. 또 IQR을 사용하여 새로운 경계를 설정하여 이상치의 검출이 가능하다는 것도 Boxplot의 큰 장점입니다. Boxplot의 단점은 바로 세부적인 데이터의 정보를 나타낼 수 없다는 것입니다. Five-number summary를 사용하기 때문에 세부 데이터의 값과 성질을 파악하기가 불가능합니다. 이는 Boxplot이 전체적인 데이터의 분포를 보여주기 때문입니다. 따라서 Boxplot과 다른 그래프를 동시에 제시하여 전체 데이터의 분포, 세부 데이터의 특징을 함께 나타내는 것이 바람직합니다. 지금까지 Boxplot에 대해 설명하였습니다. 요약하면 Boxplot은 five-number summary를 이용하여 그릴 수 있는 그래프로 빅데이터를 쉽게 한눈에 나타낼 수 있는 장점이 있으며 이상치 검출이 가능합니다. 하지만 전달하는 정보력의 한계가 있어 다른 시각화 방법이 병행되어야 합니다.