1번째 슬라이드

(간단히 데이터 리마인드의 느낌으로)

2021 빅콘테스트의 홍수ZERO부문에서 주어진 25개의 홍수 데이터를 통해 26번째 홍수의 유입량을 시간별로 예측하는 프로젝트를 수행했으며,

데이터는 홍수사상번호, 연, 월, 일, 시간 그리고 7개의 변수로 이뤄진 6개의 데이터 집단을 받았음.

결측치는 없었으며,

데이터 집단 간의 변수의 값의 차이는 거의 없었으며, 홍수사상번호 별로 시계열 데이터라는 점을 확인할 수 있었음.

홍수사상번호 별로 EDA를 수행하였음.

2번째 슬라이드

홍수의 연속적인 특성 상 과거와 미래의 데이터를 섞어봤을 때 상관 관계가 더 높을 수도 있다는 점에서 아이디어를 착안하여, Lagging을 적용.

바로 이전 시점의 변수의 데이터를 현재 시점의 데이터로 -1 시점을 당겨 옴.

각 홍수사상번호의 유입량과 변수 데이터의 상관관계가 조금 높아진 것을 heatmap을 통해 확인할 수 있었음.

또한, 히트맵에서 유입량과 수위(E지역)의 상관관계가 매우 높은 것을 확인할 수 있음.

단, 21번과 23번 홍수에서는 타 집단과는 다르게 상대적으로 낮은 수치를 보였음.

5번 홍수의 경우 수위(D지역)의 값이 모두 동일하여 결과가 나오지 않았고,

8번 집단의 경우 타 집단과 조금 다른 상관 관계를 나타낸다는 점을 확인 할 수 있었음.

3번째 슬라이드

히트맵 상에서 유입량과 상관관계가 가장 높았던 수위(E지역)을 X’ = X^2 변환을 통해 convex했던 관계를 좀 더 명확한 선형 관계로 바꾸어 줌.

주어진 데이터 프레임에 6개의 데이터 집단의 수위(E지역)의 제곱을 칼럼으로 추가하여 예측력을 높여볼 수 있는지 시도해 봄.