신한은행 3차 멘토 미팅

2021-10-28

Q1

- 데이터 처리 방법

null값과 이상치 처리를 같은 기능으로 처리해도 괜찮을까요?
(= 하나의 check box 로 둘 다 제거하는 방향도 괜찮은지)

Q2

- 이상치 분류 후 처리 방법

이상치 분류 후 삭제 or 값 대체 or 옵션 제공

Q3

- 이상치 분류 기준

"다변량 이상치 제거" 방법을 써야 되는지?
아니면 각 컬럼을 "단변량 이상치 제거" 방법으로만 하면 될지?

Q4

- 이상치 데이터 제시 방법

1. 이상치 제거 기준 후보를 옵션 제공
2. 분포에 따라서 다른 기준을 **알아서** 적용해서 결과를 보여줄지
3. 혼합방법: skewed 되어있으면 IQR적용인데 기본 IQR or 수정 IQR 쓸지는 옵션 제공

Q5

- 데이터 분량에 따른 처리

데이터 량을 충분히 가지고 있지 않다면 결과 분석을 진행하지 못하도록 하는 장치가 필요할까요?
(그렇게 적은 양의 데이터가 투입될 일이 없나요?)

Q6

- 데이터 인식

한 컬럼 내에서도 수치와 '-'가 섞여 있는 경우에는 String으로 인식하던데, 이렇게 정제되지 않은 데이터가 넘어오는 경우도 고려해야 하나요?

Q7

Pyspark

200mb- 30초 (col : 150개, rec : 200,000개)
1.2gb - 2분 30초 (col : 150개 가량, rec : 2,000,000개)
속도 부분은 개선이 어려울 것 같은데, 분석과 outlier 필터 부분에 집중해도 괜찮을지?
컬럼 수에 제한을 걸어두면 속도가 향상 될 것
> 어차피 한 화면에서 볼 수 있는 컬럼은 제한됨
> 최소 몇 개 까지는 보여줘야 좋을지 (5개 정도면 좋을것 같다?)