정규화 작업

필요한 데이터들

스나이퍼 점수 정규화를 위한 선행 데이터의 필요성

스나이퍼 점수 정규화의 목적은 체인이나 규모의 차이를 제거하고, 같은 조건 집단(코호트) 안에서 스나이퍼 원점수를 표준화 점수로 변환하는 것입니다. 이를 위해서는 각 토큰이

어떤 코호트에 속하는지, 그리고 그 코호트의 과거 점수 분포가 어떠한지 알아야 합니다.

1. `t0\_timestamp` (해당 토큰의 첫 스왑 시각)

\* 필요성: 스나이퍼는 '런칭 직후'의 매수자를 의미합니다. '런칭 직후'의 기준점을 정확히 잡기 위해서는 해당 토큰의 유동성이 처음으로 공급되어 거래가 가능해진 시점, 즉 '첫

스왑 시각'이 필요합니다. 이 시점으로부터 몇 초/몇 블록 내의 거래를 스나이퍼 활동으로 간주할 수 있습니다.

\* 활용: 스나이퍼 탐지 기준 중 '유동성 공급과 동시에 매수 발생'을 판단하는 핵심 기준이 됩니다.

2. `initial\_pool\_tvl\_usd` (t₀ 시점 초기 풀 TVL)

\* 필요성: 스나이퍼 점수 정규화의 '코호트 설계'에서 '유동성 분위(Liquidity Quartile)'를 결정하는 데 사용됩니다. 토큰의 초기 유동성 규모에 따라 스나이퍼 활동의 의미가 달라질

수 있기 때문에, 같은 규모의 토큰들끼리 비교하기 위함입니다.

\* 활용: LIQ\_Q (Liquidity Quartile)를 계산하여 cohort\_id를 생성하는 데 사용됩니다.

3. `initial\_marketcap\_usd` (t₀ 시점 시총)

\* 필요성: 스나이퍼 점수 정규화의 '코호트 설계'에서 '시총 구간(MarketCap Bin)'을 결정하는 데 사용됩니다. 시가총액 규모에 따라 스나이퍼 활동의 의미가 달라질 수 있기 때문에,

같은 규모의 토큰들끼리 비교하기 위함입니다.

\* 활용: MCAP\_BIN (MarketCap Bin)을 계산하여 cohort\_id를 생성하는 데 사용됩니다.

4. `raw\_sniper\_score\_0\_100` 표본 데이터 (과거 데이터베이스)

\* 필요성: 스나이퍼 점수 정규화의 핵심인 '표준화 점수(z\_vdw, robust\_z)'를 계산하기 위한 '베이스라인(Baseline)'을 구축하는 데 사용됩니다. 특정 코호트(예: ETH-Q2-Small)에

속하는 토큰들의 과거 스나이퍼 점수 분포를 알아야, 현재 토큰의 점수가 그 코호트 내에서 어느 정도 위치인지(퍼센타일)를 판단할 수 있습니다.

\* 활용: 각 코호트별로 median, IQR 등을 계산하여 z\_vdw와 robust\_z를 산출하는 데 사용됩니다. 최소 30개 이상의 표본이 필요하며, 이를 위해서는 파이프라인이 분석한 과거

데이터를 저장할 데이터베이스(DB)가 필수적입니다.

---

이 데이터들은 스나이퍼 점수 정규화의 정확성과 객관성을 확보하는 데 필수적이지만, 현재 파이프라인의 데이터 수집 범위를 넘어서는 부분입니다. 특히 '초기 풀 TVL'과 '초기

시총'은 알케미 API로 조회하기 매우 까다롭고, '과거 스나이퍼 점수 표본 데이터'는 별도의 데이터베이스 구축이 필요합니다.

1. Snipers
2. 절댓값으로 계산했을 때 문제: 네트워크 간의 상이함 + 규모가 다름

* 예) ETH 메인넷 런치는 가스비 자체가 높다
* 예) 유동성 100만불과 10만불은 스나이퍼 움직이는 단위 자체가 다름
* 절댓값 크기 자체로 비교하면 왜곡이 발생
* 정규화 작업이 필요함
* Cohort 분리 작업(같은 집단/기준끼리 묶어서 표준화)

1. 분리 기준
2. 네트워크&체인: ETH, BSC, and Solana
3. 풀 규모: 런치 시점 초기 풀 TVL

* 풀이 얕으면 소액 스나도 영향이 크고 그 반대도 성립함
* Q1~Q4로 구분

1. 토큰 시총/ 발행량 구간: 시총이 크면 스나이퍼 물량 자체가 커짐

* Micro/Small/Mid 이렇게 세 개로 구분
* 기본 모델. 런칭 초기에 샘플 부족할 때 이 세개만을 고려.

1. DEX 타입: ETH/BSC= v2/v3, SOL=단일 AMM

* 수수료 티어, 가격곡선 구조가 달라서 동일한 패턴도 체감효과가 다름
* 데이터가 안정적으로 쌓이면 추가 (약 72코호트)

1. 시간 레짐: 밈코인 런치 환경은 빠르게 변하므로 오래된 데이터와 섞으면 왜곡 발생

* 풀옵. 고도화 단계에서 반영. (~120코호트)