### lof

가장 인접한 k개의 이웃을 기준으로 각 점의 밀도를 계산하고, 인접한 이웃과 상대적인 밀도의 비율을 계산하여 LOF (Local Outlier Factor) 지수를 계산합니다.

#### 문법

lof [k=INT] FIELD, ... [by GRP\_FIELD, ...]

필수 매개변수

**FIELD, ...**

정수, 실수, 날짜와 같은 숫자로 구성된 값을 포함한 필드 목록. 구분자로 쉼표(,)를 사용합니다.

선택 매개변수

**eps=DOUBLE**

데이터 간 최소 거리 조정계수(기본값: 0.00001). 데이터 간 거리의 합을 나눈 값이 무한으로 발산하지 않도록 조정하기 위해 사용합니다.

**k=INT**

계산에 사용할 이웃 노드 개수(기본값: 10).

**[by GRP\_FIELD\_1, ...]**

LOF 지수 계산에 사용할 그룹핑 필드 목록. 구분자로 쉼표(,)를 사용합니다. 이 옵션은 FIELD, ... 뒤에 지정해야 합니다.

by 절을 사용해 그룹별 스코어링을 계산하려면 각 그룹의 레코드 수는 이웃 노드 수(k=INT으로 지정한 값)보다 많아야 합니다. 이웃 노드 수보다 그룹의 레코드 수가 적으면 모든 점이 하나의 군집으로 잡히기 때문에 LOF 지수(**\_lof** 필드)가 의도한대로 계산되지 않습니다.

#### 설명

각 레코드마다 \_lof 필드에 LOF 지수를 계산하며, 이 값은 다음과 같이 분류할 수 있습니다:

* 값이 1보다 클 때(LOF(k) > 1): 군집의 바깥쪽에 위치합니다. 1보다 클수록 이상치(anomaly)일 가능성이 높습니다.
* 값이 1의 근사값일 때(LOF(k) ≈ 1): 군집의 경계에 위치합니다.
* 값이 1보다 작을 때(LOF(k) < 1): 군집의 내부에 위치합니다.

#### 사용 예

**sepal\_length**와 **sepal\_width** 필드 값을 기준으로 이상치를 계산합니다(다운로드: <https://raw.githubusercontent.com/illinois-cse/data-fa14/gh-pages/data/iris.csv>).

wget url="https://raw.githubusercontent.com/illinois-cse/data-fa14/gh-pages/data/iris.csv" | eval line = split(line, "\n") | explode line | split sep="," sepal\_length,sepal\_width,petal\_length,petal\_width,species| eval sepal\_length = double(sepal\_length), sepal\_width = double(sepal\_width)| lof sepal\_length, sepal\_width| search \_lof > 2