Calculating the Outliers

Coffeebeans Data Engineer Challenge

Calculation:

A week is classified as an outlier when the total votes for the week deviate from the average votes per week for the complete dataset by more than 20%. For the avoidance of doubt, *please use the following formula*:

Say the mean votes is given by $\bar{x}$ and this specific week's votes is given by x_i.
We want to know when x_i differs from $\bar{x}$ by more than 20%.
When this is true, then the ratio $\frac{x_i}{\bar{x}}$ must be further from 1 by more than 0.2, i.e.:

$$\left|1 - \frac{x_i}{\bar{x}}\right| > 0.2$$

We want this outlier calculation's output to be stored in the view called outlier_weeks. The data should be sorted in the view by year and week number, with the earliest week first.

Test Data

Given the following test data:

{"Id":"1","PostId":"1","VoteTypeId":"2","CreationDate":"2022-01-02T00:00:00.000"}

{"Id":"2","PostId":"1","VoteTypeId":"2","CreationDate":"2022-01-09T00:00:00.000"}

{"Id":"4","PostId":"1","VoteTypeId":"2","CreationDate":"2022-01-09T00:00:00.000"}

{"Id":"5","PostId":"1","VoteTypeId":"2","CreationDate":"2022-01-09T00:00:00.000"}

{"Id":"6","PostId":"5","VoteTypeId":"3","CreationDate":"2022-01-16T00:00:00.000"}

{"Id":"7","PostId":"3","VoteTypeId":"2","CreationDate":"2022-01-16T00:00:00.000"}

{"Id":"8","PostId":"4","VoteTypeId":"2","CreationDate":"2022-01-16T00:00:00.000"}

{"Id":"9","PostId":"2","VoteTypeId":"2","CreationDate":"2022-01-23T00:00:00.000"}

{"Id":"10","PostId":"2","VoteTypeId":"2","CreationDate":"2022-01-23T00:00:00.000"}

{"Id":"11","PostId":"1","VoteTypeId":"2","CreationDate":"2022-01-30T00:00:00.000"}

{"Id":"12","PostId":"5","VoteTypeId":"2","CreationDate":"2022-01-30T00:00:00.000"}

{"Id":"13","PostId":"8","VoteTypeId":"2","CreationDate":"2022-02-06T00:00:00.000"}

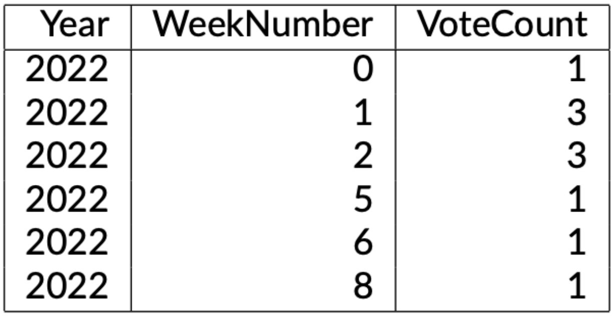
{"Id":"14","PostId":"13","VoteTypeId":"3","CreationDate":"2022-02-13T00:00:00.000"}

{"Id":"15","PostId":"13","VoteTypeId":"3","CreationDate":"2022-02-20T00:00:00.000"}

{"Id":"16","PostId":"11","VoteTypeId":"2","CreationDate":"2022-02-20T00:00:00.000"}

{"Id":"17","PostId":"3","VoteTypeId":"3","CreationDate":"2022-02-27T00:00:00.000"}

You should have the following in your outlier_weeks view:



Year	WeekNumber	VoteCount
2022	0	1
2022	1	3
2022	2	3
2022	5	1
2022	6	1
2022	8	1

Note that we strongly encourage you to use this data as a test case to ensure that you have the correct calculation!