Proposed metrics for classifying businesses as “popular” and “successful”

All metrices Information

Before classifying a business as “popular” or “successful”, there are six metrics used for classification. (1) Business matrix: it has a business unique id and contains business names, addresses, cities, states, zip codes, star scores, review numbers, number of total check-ins, and average star rating for each business. (2) Attributes matrix: it refers to the business matrix and contains the presence or absence of objects (e.g., toilets, parking lots, etc.) included in each business. (3) Categories matrix: it refers to the business matrix and contains categories of each business. (4) Check-in matrix: it refers to the business matrix and includes check-in times, days of the week, and the number of business checks-in. (5) Users matrix: it has a user's unique id and includes the number of reviews written by users, the average rating of all reviews, the number of funny votes, the number of useful votes, and the number of cool votes. (6) Review matrix: it refers to the business and user matrices and includes star scores, dates, number of funny votes, number of useful votes, and number of cool votes.

Popular businesses

The stars, total number of reviews, total check-in, and review rating of the business matrix were judged to be good objects for classifying popular businesses, but I don't think the rest of the objects except the total check-in are suitable for the term "popular". For example, in the case of McDonald's, stars, total number of reviews, and review ratings are lower than other companies, but more total check-in than others (people more visit than other companies). This means that the company "McDonald" is already widely known to the public. Therefore, to set the criteria for classifying popular businesses in each zip code, it is judged that it is better to list the businesses with the highest number of check-ins for each category in the selected zip code. First, you can select popular businesses for each region based on the results of obtaining the total number of check-ins from the check-in metric and then updating them to the 'numcheckins' object of the business metric (the larger the check-in total, the more popular the business is).

The INSERT query used to process data and derive results:

```
SELECT cs.cname, bs.bname, ROUND(bs.stars,1), bs.numcheckins, bs.reviewrating,
bs.reviewcount
FROM business as bs, categories as cs
WHERE cs.bid = bs.bid AND bs.zip = ' ' AND
(cs.cname, bs.numcheckins) IN (SELECT cname, MAX(numcheckins)
                               FROM business, categories
                               WHERE business.bid = categories.bid AND
                                     business.zip = ' '
                               GROUP BY cname)
GROUP BY cname
ORDER BY cs.cname;
```

This query outputs category names, business names, star scores, total check-in numbers, review ratings, and total number of reviews. Businesses with the highest total check-ins among the businesses by category belonging to each zip code are grouped by category, and only one business with the highest total check-ins for each category is extracted. After that, sort by the categories name (alphabetical order.).

Successful businesses

Successful business refers to a business that has long been loyal to volunteer work in the community. The postal code selected to set the criteria for classification of successful industries calculates the average number of check-ins in the same category, and if the total number of check-ins is greater than this average number of check-ins, it is determined as a successful industry.

The INSERT query used to process data and derive results:

```
SELECT bs.bname, bs.address, ROUND(bs.stars,1), bs.reviewcount, bs.numcheckins
FROM business as bs, categories as cs
WHERE cs.bid = bs.bid AND
bs.zip = ' ' AND
(bs.numcheckins) > (SELECT AVG(numcheckins)
                                    FROM business, categories
                                    WHERE business.bid = categories.bid AND
                                    business.zip = ' ' AND
                                            categories.cname = cs.cname
                                    GROUP BY cname)
GROUP BY bs.bname, bs.address, bs.stars, bs.reviewcount, bs.numcheckins
ORDER BY bs.numcheckins DESC;
```

This query outputs the name, address, star, total number of reviews, and total number of check-in. Here, through subqueries, the total check-in average for each business for each zip code is obtained. After that, only businesses with a total number of check-ins higher than the check-ins average for each business for each zip code are extracted. Then sort based on the most total check-ins.