**TASK 4**

**Try to understand this dataset and how fields are generated by Yelp. How are stars (business table) calculated and what are they composed of.**

Za izradu ovog taska sam koristila *review.json* dataset. U njemu se nalaze podaci svakog ostavljenog review-a, kao npr. ocjena, teks, id korisnika, biznisa. Zbog velike količine podataka (više od 5GB), korišteno je prvih 100 000 redova. Statistički gledano, to bi trebao biti dovoljan set podataka da možemo, sa određenom sigurnošću, odrediti način računanja konačne ocjene biznisa.

Za početak je .json tip konvertovan u .csv format, na isti način kao u prvom zadatku, te je kreirana tabela review koja čuva sve te podatke.

Generisanje tabele table *business\_id\_review\_count* koja sadrži dvije kolone, *business\_id* i *count*, a one prestavljaju sumu svih review-a za svaki pojedini biznis koji imamo u datasetu *review.*

select distinct business\_id, count(business\_id)

into table business\_id\_review\_count

from review

group by business\_id

Ukoliko napišemo ovaj query:

select business\_id\_review\_count.business\_id, ybusiness.review\_count,

business\_id\_review\_count.count, stars, avg\_stars.avg

from ybusiness, business\_id\_review\_count, avg\_stars

where business\_id\_review\_count.business\_id = ybusiness.business\_id and

ybusiness.business\_id = avg\_stars.business\_id and

business\_id\_review\_count.count > 5 and

100 \* business\_id\_review\_count.count / review\_count > 20

Rezultat je 912 različitih biznisa koje uzimamo kao relevntne. Uslovi su da su ponavljaju u review datasetu više od 5 puta i da je postotak prisutnih u odnosu na ukupni broj, korišten za određivanje konačne ocjene, veći od 20%.

Nakon toga provjeravamo koliko je stvarnih ocjena približno jednako procijenjenoj ocjeni.

select business\_id\_review\_count.business\_id, ybusiness.review\_count,

business\_id\_review\_count.count, stars, avg\_stars.avg

from ybusiness, business\_id\_review\_count, avg\_stars

where business\_id\_review\_count.business\_id = ybusiness.business\_id and

ybusiness.business\_id = avg\_stars.business\_id and

business\_id\_review\_count.count > 5 and

100 \* business\_id\_review\_count.count / review\_count > 20 and

stars between avg\_stars.avg - 0.5 and avg\_stars.avg + 0.5

Kao rezultat sam dobila da je to 765.

Dakle, na osnovu obrađivanja 5% od ukupnog broja podataka, možemo, sa sigurnošću od 84% zaključiti da se zvijezde na Yelpu dodjeljuju jednostavno određujući aritmetičku sumu svih ocijena koje je taj biznis dobio.

Neki od tih podataka su prikazani u sljedećoj tabeli.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **business\_id** | **review\_count** | **count** | **stars** | **avg** |
| "n8hRciFiLRm1uuX4iBjJhg" | 17 | 6 | 4.5 | 4.33333333333333 |
| "iHdg2E\_z6CyOurYTm3WHag" | 26 | 7 | 5 | 5 |
| "rmOZm59ahqMJlfGhNg0thw" | 56 | 16 | 3.5 | 3.4375 |
| "v9VCOZbkQJxdw4-iY1ZXZQ" | 23 | 6 | 4 | 3.5 |
| "lT-fvWPR-ju0-kNVQgs4Ag" | 27 | 7 | 4.5 | 4.14285714285714 |
| "dGL86QPuTDZBu0P7XvrIKw" | 120 | 28 | 3 | 3.17857142857143 |
| "06X34OSshQJPDiyUDHtS0A" | 34 | 8 | 4 | 4.125 |
| "XmgbQmB8SiFcWVWE1xU3fw" | 29 | 8 | 3.5 | 3.625 |
| "DxiT0Jc2iAjtml-XXBhb1A" | 37 | 10 | 4 | 4.3 |
| "q7f\_lCOKf40idrVSzmoYug" | 33 | 7 | 3 | 2.57142857142857 |
| "t\_rnyBYcjuCa5m-HWU0ZwQ" | 25 | 8 | 3.5 | 3 |
| "dM8i3hklYuGE2xiVsQZ4Yg" | 71 | 17 | 3.5 | 3.70588235294118 |
| "MXTULCzHS5r\_9SkxgvntVw" | 33 | 8 | 3 | 2.5 |
| "c3wb5TZog6ngKMTlMHUaZQ" | 12 | 6 | 3.5 | 3.16666666666667 |
| "ZSONEY8fvrMHyK7vzK4yPA" | 19 | 6 | 2 | 1.83333333333333 |