PROFILING AND ANALYZING THE YELP DATASET: PART 2: INFERENCES AND ANALYSIS

For this part of your analysis, you are going to choose the type of analysis you want to conduct on the Yelp dataset and are going to prepare the data for analysis.

i. Indicate the type of analysis you chose to do:

I wanted to see if the stars and check-in count have any correlation with negative words used in the actual review.

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:

I will be using review and checkin data for this analysis as they contain the star rating, checkin count for businesses.

For this analysis, I have used positive words, such as, best, good, love, and like and for negative reviews, hate, terrible, and bad were used.

CODES FOR NEGATIVE SENTIMENT:

```
SELECT DISTINCT(r.business_id), r.stars, ch.count
FROM  checkin ch INNER JOIN review r ON r.business_id = ch.business_id
WHERE  text like '%terrible%'
ORDER BY r.stars DESC
```

```
+------------------------+-------+-------+
| business_id            | stars | count |
+------------------------+-------+-------+
| 3C5cDapsxHKckHn5-cmCzg |     2 |     3 |
| 3C5cDapsxHKckHn5-cmCzg |     2 |     2 |
| 3C5cDapsxHKckHn5-cmCzg |     2 |     1 |
| 3C5cDapsxHKckHn5-cmCzg |     2 |     4 |
+------------------------+-------+-------+
```

```
SELECT DISTINCT(r.business_id), r.stars, ch.count
FROM  checkin ch INNER JOIN review r ON r.business_id = ch.business_id
WHERE  text like '%hate%' OR text like '%bad%' OR text like '%terrible%'
ORDER BY r.stars DESC
```

```
+------------------------+-------+-------+
| business_id            | stars | count |
+------------------------+-------+-------+
| 3C5cDapsxHKckHn5-cmCzg |     2 |     3 |
| 3C5cDapsxHKckHn5-cmCzg |     2 |     2 |
| 3C5cDapsxHKckHn5-cmCzg |     2 |     1 |
| 3C5cDapsxHKckHn5-cmCzg |     2 |     4 |
+------------------------+-------+-------+
```

AS shown above, there are only four matching records were found which have the word "terrible" in their reviews. However, none of the matching records has the other three negative words. As expected, these businesses have low rating (2/5) and low checkin count

CODES FOR POSITIVE SENTIMENT:

```
SELECT DISTINCT(r.business_id), r.stars, ch.count
FROM  checkin ch INNER JOIN review r ON r.business_id = ch.business_id
WHERE  text like '%best%'
ORDER BY r.stars DESC
```

business_id	stars	count
qchCmKqTY46vKS1Y8zXdQQ	5	1
u_1EIg-c3d8XYfrmcdHX3w	4	8
u_1EIg-c3d8XYfrmcdHX3w	4	2
u_1EIg-c3d8XYfrmcdHX3w	4	1
u_1EIg-c3d8XYfrmcdHX3w	4	15
u_1EIg-c3d8XYfrmcdHX3w	4	4
u_1EIg-c3d8XYfrmcdHX3w	4	9
u_1EIg-c3d8XYfrmcdHX3w	4	3
u_1EIg-c3d8XYfrmcdHX3w	4	5
u_1EIg-c3d8XYfrmcdHX3w	4	6
u_1EIg-c3d8XYfrmcdHX3w	4	7
u_1EIg-c3d8XYfrmcdHX3w	4	16

```
SELECT DISTINCT(r.business_id), r.stars, ch.count
FROM  checkin ch INNER JOIN review r ON r.business_id = ch.business_id
WHERE  text like '%love%'
ORDER BY r.stars DESC
```

business_id	stars	count
6RxtO_MxnLwZ7D7iYDE8Jg	5	1
6RxtO_MxnLwZ7D7iYDE8Jg	5	3
6RxtO_MxnLwZ7D7iYDE8Jg	5	2
6RxtO_MxnLwZ7D7iYDE8Jg	5	4
6RxtO_MxnLwZ7D7iYDE8Jg	5	5
6RxtO_MxnLwZ7D7iYDE8Jg	5	8
lxsXOZNKoKPwXgkigdWPng	4	3
lxsXOZNKoKPwXgkigdWPng	4	1
lxsXOZNKoKPwXgkigdWPng	4	2

```
SELECT DISTINCT(r.business_id), r.stars, ch.count
FROM  checkin ch INNER JOIN review r ON r.business_id = ch.business_id
WHERE  text like '%like%'
ORDER BY r.stars DESC
```

business_id	stars	count
qchCmKqTY46vKS1Y8zXdQQ	5	1

```
SELECT DISTINCT(r.business_id), r.stars, ch.count
FROM  checkin ch INNER JOIN review r ON r.business_id = ch.business_id
WHERE  text like '%good%'
ORDER BY r.stars DESC
```

business_id	stars	count
ebgu1HjoE5S45U7oKyq4sw	5	2
ebgu1HjoE5S45U7oKyq4sw	5	1
ebgu1HjoE5S45U7oKyq4sw	5	5
ebgu1HjoE5S45U7oKyq4sw	5	3
ebgu1HjoE5S45U7oKyq4sw	5	4
8C65DwwnazwYB2z9naHyaw	5	2
8C65DwwnazwYB2z9naHyaw	5	1
8C65DwwnazwYB2z9naHyaw	5	3
8C65DwwnazwYB2z9naHyaw	5	4
u_1EIg-c3d8XYfrmcdHX3w	4	8
u_1EIg-c3d8XYfrmcdHX3w	4	2
u_1EIg-c3d8XYfrmcdHX3w	4	1
u_1EIg-c3d8XYfrmcdHX3w	4	15
u_1EIg-c3d8XYfrmcdHX3w	4	4
u_1EIg-c3d8XYfrmcdHX3w	4	9
u_1EIg-c3d8XYfrmcdHX3w	4	3
u_1EIg-c3d8XYfrmcdHX3w	4	5
u_1EIg-c3d8XYfrmcdHX3w	4	6
u_1EIg-c3d8XYfrmcdHX3w	4	7
u_1EIg-c3d8XYfrmcdHX3w	4	16
5MXId9GCUlYgokbvqayzVg	4	1
3C5cDapsxHKckHn5-cmCzg	2	3
3C5cDapsxHKckHn5-cmCzg	2	2
3C5cDapsxHKckHn5-cmCzg	2	1
3C5cDapsxHKckHn5-cmCzg	2	4

As shown above, there were 35 matching records with positive sentiments. All of them have high star rating (4-5) except the last four with "good" in their reviews. However, the check in count does not show any direct correlation with reviews or the ratings. Although there are few businesses in the positive sentiment records with high check in counts most of them small number of check ins similar to the "terrible" ones. Although there are many other factors that need to be considered which are beyond the scope of this analysis, such as, the context of those words, this analysis shows that most often there is a correlation between positive review and star rating but no correlation between rating or reviews and check in count. Therefore based on this data it is difficult to conclude whether more people visit businesses with higher ratings and positive reviews only.