TEAM BLUE YELP DATA EXPLORATION & ANALYSIS

August 6, 2022

HYPOTHESIS AND QUESTION

We believe that the top 100 users, based on highest number of useful votes, will have a lower average star rating and higher than average review count than the bottom 100 users in the data set.

Did the average ratings by the top 100 users change pre- and post-covid?

STEPS TO DIG INTO THE DATA

1. Find top 100 users based on highest number of useful votes
2. Find the average star rating and average number of reviews for these 100
3. Find the average star rating and average number of reviews for the 100 least useful raters (assume that least useful raters have >1000 reviews, >0 useful votes)
4. Review average ratings for the top 100 - before covid (1/1/2020)
5. Review average ratings for all but the top 100 - before covid
6. Review average ratings for the top 100 - after covid (1/1/2020)
7. Review average ratings for all but the top 100 - after covid

ASSUMPTIONS

1. COVID begins January 1, 2020
2. Top 100 Most Useful Users defined as:
   - Most useful votes from other Yelpers
3. Bottom 100/Least Useful Users defined as:
   - Greater than 1,000 reviews
   - At least one useful vote from other Yelpers
   - Didn't want to include the significant number of reviews/users with reviews who had 0 useful votes since that could confuse the data subset and results
4. We understand that the total number of reviews included every review the user has created even though not every one of those reviews were in the data. We believe this is due to the Yelp data only being comprised of certain cities' reviews on a by-business basis, but the user data being representative of all their respective reviews.

CODE FOR PART 1 & ANALYSIS

1. **Find top 100 users based on highest number of useful votes**
   - Code:

```
USE yelp;

SELECT DISTINCT(user_id), u.name, u.review_count, u.average_stars, u.useful_votes
  FROM user u
  INNER JOIN review r
  USING (user_id)
  ORDER BY useful_votes DESC
  LIMIT 100;
```

   - Analysis:
     - The Top 100 users, based on highest number of useful votes, had 259,132 total reviews
     - Their reviews accounted for 6,153,821 useful votes from other Yelp users

CODE FOR PART 2 & ANALYSIS

1. **Find the average star rating and average number of reviews for these 100**
   - Code:

```
USE yelp;

SELECT DISTINCT(user_id), u.name, u.review_count, u.average_stars, u.useful_votes
  FROM user u
  INNER JOIN review r
  USING (user_id)
  ORDER BY useful_votes DESC
  LIMIT 100;
```

   - Analysis:
     - The Top 100 users' average star rating: 3.95
     - The Top 100 users' average number of reviews: 2,591

CODE FOR PART 3 & ANALYSIS

1. **Find the average star rating and average number of reviews for the 100 least useful raters (>1000 reviews, >0 useful votes)**
   - Code:

     ```
     SELECT DISTINCT(user_id), u.name, u.review_count, u.average_stars, u.useful_votes
       FROM user u
       INNER JOIN review r
       USING (user_id)
       WHERE u.useful_votes >'0' AND u.review_count >='1000'
       ORDER BY useful_votes ASC
       LIMIT 100;
     ```

   - Analysis:
     - The Top 100 users' average star rating: 3.75
     - The Top 100 users' average number of reviews: 1,233

KEY INSIGHT:
TOP 100 VS. BOTTOM 100

	Top 100 Most Useful	Least Useful 100	% Change Top vs. Bottom
Average Rating	3.95	3.75	+5.33%
Average Number of Reviews	2,591	1,233	+110.14%

CODE FOR PART 4 & ANALYSIS

1. **Review average ratings for the top 100 - before covid**
   - Code:

```
SELECT DISTINCT(user_id), name, review_count, average_stars, useful_votes, review_date, stars
 FROM top_100_users
 LEFT JOIN review r
 USING (user_id)
 WHERE review_date < '2020-01-01'
 ORDER BY useful_votes DESC;
```

   - Analysis:
     - Average star rating (average of averages): 3.8186

CODE FOR PART 5 & ANALYSIS

1. **Review average ratings for all but the top 100 - before covid**
   - Code:

```
SELECT DISTINCT(user_id), name, review_count, average_stars, useful_votes, stars
FROM bottom_100_users
INNER JOIN review r
USING (user_id)
WHERE useful_votes >'0' AND review_count >='1000' AND review_date < '2020-01-01'
ORDER BY useful_votes ASC;
```

   - Analysis:
     - Average star rating (average of averages): 3.4294

CODE FOR PART 6 & ANALYSIS

1. **Review average ratings for the top 100 – post-covid (1/1/2020)**
   - Code:

     ```
     SELECT DISTINCT(user_id), name, review_count, average_stars, useful_votes, review_date, stars
      FROM top_100_users
      LEFT JOIN review r
      USING (user_id)
      WHERE review_date >= '2020-01-01'
      ORDER BY useful_votes DESC;
     ```

   - Analysis:
     - Average star rating (average of averages): 4.0586

CODE FOR PART 7 & ANALYSIS

1. **Review average ratings for all but the top 100 – post-covid (1/1/2020)**
   - Code:

```
SELECT DISTINCT(user_id), name, review_count, average_stars, useful_votes, stars
FROM bottom_100_users
INNER JOIN review r
USING (user_id)
WHERE useful_votes >'0' AND review_count >='1000' AND review_date >= '2020-01-01'
ORDER BY useful_votes ASC;
```

   - Analysis:
     - Average star rating (average of averages): 3.6742

KEY INSIGHT:
PRE VS. POST-COVID (1/1/2020)

	Top 100 Most Useful	Least Useful 100	% Change Top vs. Bottom
Pre-Covid Avg Rating	3.8186	3.4294	**+11.35%**
Post-Covid Avg Rating	4.0568	3.6742	**+10.43%**
% Change Pre vs. Post	**+6.238%**	**+7.138%**	

THANK YOU

Team Blue