RBDA Spring 2023 Project :

Yelp Review Analysis

Team Members:
Se Jin Lee (sl9339)
Ajeeta Asthana (aa9381)
Rachid Tak Tak (rt2504)
Reuben Abraham (rca9751)

Abstract

- Yelp is a website that publishes crowd-sourced reviews about businesses.
- This dataset is a publicly accessible repository of Yelp businesses, users, reviews and tips. It was originally put together for the Yelp Dataset Challenge which was a competition for students to conduct analysis on Yelp data and share their discoveries.
- Each individual dataset was originally in json form.
- Our datasets were loaded onto NYU's Dataproc cluster where we used Hadoop MapReduce for our data-cleaning and Hive+Trino for our data-analysis.

Users, Benefits and Importance

- Our analysis will help both the end-consumer, as well as the business owner.
- Our end-consumer can understand which are the best businesses grouped by geography and category.
- For Yelp and the businesses owners, our reports will help catch anomalies in reviews and track active reviewers.
- This analysis is important because it helps consumers identify the best fix for their problem- and for businesses to better serve their customers. Data from our analysis can be used to feed recommendation engines for Yelp users.

Goodness

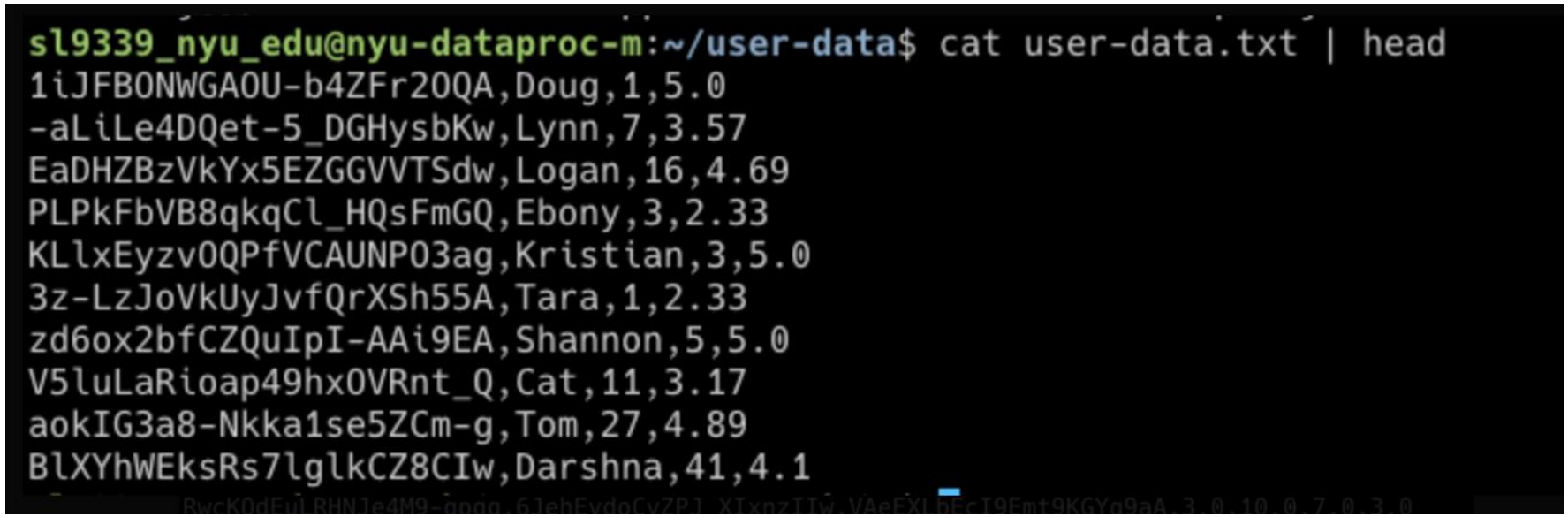
- Review table has user_id as a foreign key. Joining the two tables will provide more in-depth analysis about reviews written by a particular group of users
- Review table has business_id as a foreign key. Joining review, user and business tables made sense to find more cross data insights like users and sentiment analysis of the caetgories of businesses they are reviewing.

Data Sources

- User Data:
  - All information pertaining to a user (reviewer).
  - 3.36 GB
- Review:
  -Contains review text data including the user that wrote the review, and the business the review is written for.
  - 5.34 GB
- Business:
  -Includes information about different businesses including location, attributes, and which category it belongs to.
  - 118.86 MB
- Tips:
  - Includes tips written by users about a business. They are shorted than reviews and convey quick suggestions.
  - 180 MB

Data Sample

- User Data:



Data Sample

- Review Data:

```
sl9339_nyu_edu@nyu-dataproc-m:~/user-data$ cat review-data.txt | head
RwcKOdEuLRHNJe4M9-qpqg,6JehEvdoCvZPJ_XIxnzIIw,VAeEXLbEcI9Emt9KGYq9aA,3.0,10.0,7.0,3.0
i-I4ZOhoX70Nw5H0FwrQUA,YwAMC-jvZ1fvEUum6QkEkw,Rr9kKArrMhSLVE9a53q-aA,5.0,1.0,0.0,0.0
YNfNhgZlaaCO5Q_YJR4rEw,mm6E4FbCMwJmb7kPDZ5v2Q,R1khUUxidqfaJmcpmGd4aw,4.0,1.0,0.0,0.0
shTPgbgdwTHSuU67mGCmZQ,Zo0th2m8Ez4gLSbHftiQvg,2vLksaMmSEcGbjI5gywpZA,5.0,2.0,2.0,1.0
H0RIamZu0B0Ei0P4aeh3sQ,qskILQ3k0I_qcCMI-k6_QQ,jals67o91gcrD4DC81Vk6w,5.0,1.0,1.0,2.0
YVX1Wsa4LYxjvFwuHBb_gA,RKPkxOYQlM0BjhM-H6_vAw,X4mouE_cMiwbfyCPZ_K-FA,4.0,3.0,2.0,0.0
zHZ-A1qyKDEgyZMDaD--wg,_XVdmFWSgTN6YlojUxixTA,6WaI-IN8ql0xpEKlb4q8tg,5.0,1.0,0.0,0.0
wD5ZWao_vjyT2h4xmGam8Q,7L7GL5Pi2cf8mbm2Dpw4zw,e_E-jq9mwm7wk75k7Yi-Xw,5.0,1.0,1.0,0.0
Sm8-QDsuQfik-QuhRYT5bw,QIXYkyAbTvgePUOH0-cRFg,Bh4b8wJRR_ggH7JvpCO7CQ,2.0,0.0,0.0,0.0
7NgXAuTFiJHYbuepOPwU0w,x1QLCwZGFAjxRRw4EHc3-g,1_BVWDzi5cVqWxNe9bOMMQ,5.0,1.0,1.0,0.0
```

Data Sample

- Business Data:

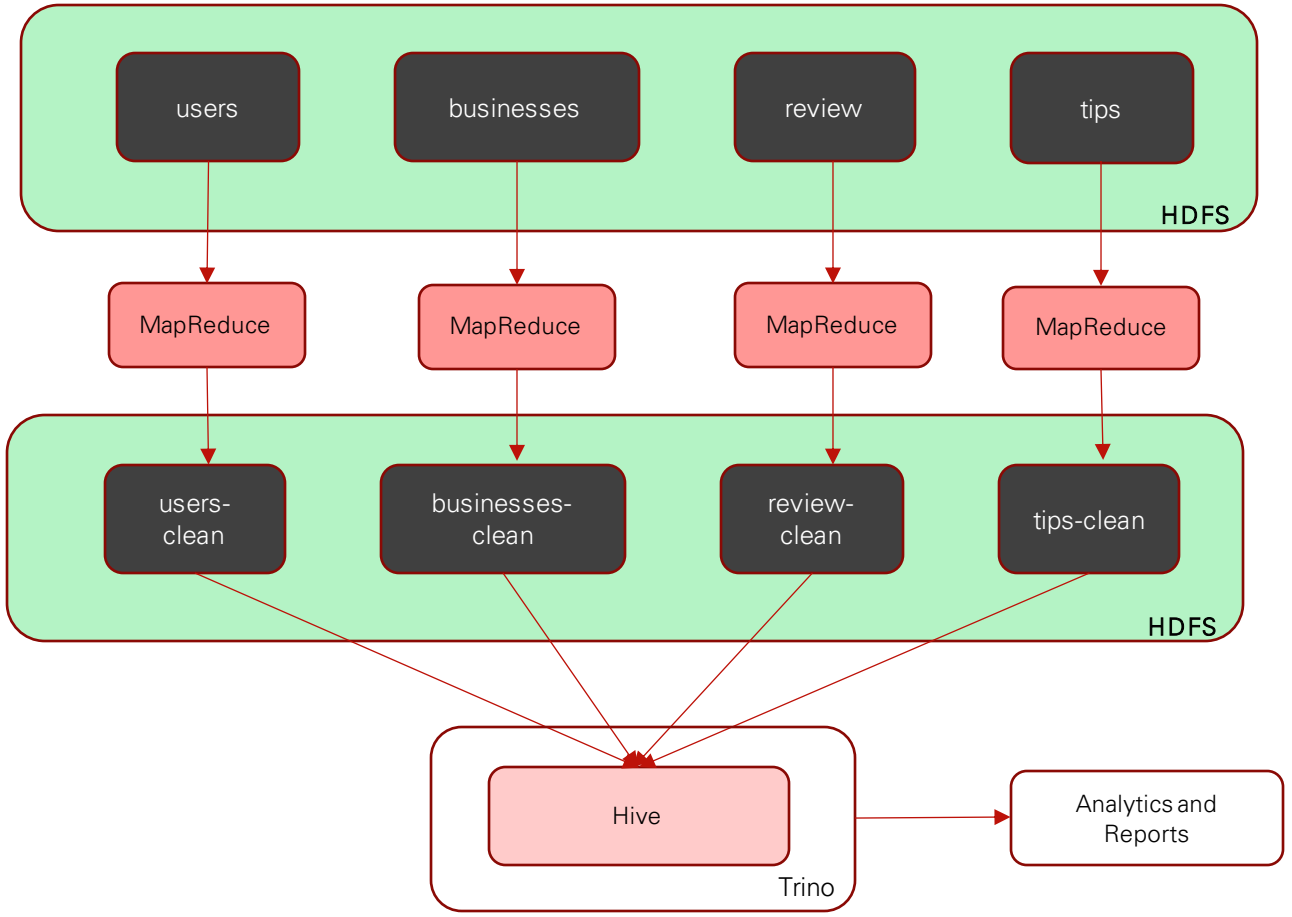
```
Pns214eNsfO8kk83dixA6A,Abby Rappoport; LAC; CMQ,1616 Chapala St; Ste 2,Santa Barbara,CA,93101,34.4266787,-119.7111968,5.0,7,Doctors; Traditional Chinese Medicine; Naturopathic/Holistic; Acupun
cture; Health & Medical; Nutritionists
mpf3x-BjTdTEA3yCZrAYPw,The UPS Store,87 Grasso Plaza Shopping Center,Affton,MO,63123,38.551126,-90.335695,3.0,15,Shipping Centers; Local Services; Notaries; Mailbox Centers; Printing Services
tUFrWirKiKi_TAnsVWINQQ,Target,5255 E Broadway Blvd,Tucson,AZ,85711,32.223236,-110.880452,3.5,22,Department Stores; Shopping; Fashion; Home & Garden; Electronics; Furniture Stores
MTSW4McQd7CbVtyjqoe9mw,St Honore Pastries,935 Race St,Philadelphia,PA,19107,39.9555052,-75.1555641,4.0,80,Restaurants; Food; Bubble Tea; Coffee & Tea; Bakeries
mWMc6_wTdE0EUBKIGXDVfA,Perkiomen Valley Brewery,101 Walnut St,Green Lane,PA,18054,40.3381827,-75.4716585,4.5,13,Brewpubs; Breweries; Food
CF33F8-E6oudUQ46HnavjQ,Sonic Drive-In,615 S Main St,Ashland City,TN,37015,36.269593,-87.058943,2.0,6,Burgers; Fast Food; Sandwiches; Food; Ice Cream & Frozen Yogurt; Restaurants
n_0UpQx1hsNbnPUSlodU8w,Famous Footwear,8522 Eager Road; Dierbergs Brentwood Point,Brentwood,MO,63144,38.627695,-90.340465,2.5,13,Sporting Goods; Fashion; Shoe Stores; Shopping; Sports Wear; Ac
cessories
qkRM_2X51Yqxk3btlwAQIg,Temple Beth-El,400 Pasadena Ave S,St. Petersburg,FL,33707,27.76659,-82.732983,3.5,5,Synagogues; Religious Organizations
k0hlBqXX-Bt0vf1op7Jr1w,Tsevi's Pub And Grill,8025 Mackenzie Rd,Affton,MO,63123,38.5651648,-90.3210868,3.0,19,Pubs; Restaurants; Italian; Bars; American (Traditional); Nightlife; Greek
bBDDEgkFA1Otx9Lfe7BZUQ,Sonic Drive-In,2312 Dickerson Pike,Nashville,TN,37207,36.2081024,-86.7681696,1.5,10,Ice Cream & Frozen Yogurt; Fast Food; Burgers; Restaurants; Food
UJsufbvfyfONHeWdvAHKjA,Marshalls,21705 Village Lakes Sc Dr,Land O' Lakes,FL,34639,28.1904587953,-82.4573802199,3.5,6,Department Stores; Shopping; Fashion
eEOYSgkmpB90uNA7lDOMRA,Vietnamese Food Truck,,Tampa Bay,FL,33602,27.9552692,-82.4563199,4.0,10,Vietnamese; Food; Restaurants; Food Trucks
il_Ro8jwPlHresjw9EGmBg,Denny's,8901 US 31 S,Indianapolis,IN,46227,39.6371332838,-86.127217412,2.5,28,American (Traditional); Restaurants; Diners; Breakfast & Brunch
jaxMSoInw8Poo3XeMJt8lQ,Adams Dental,15 N Missouri Ave,Clearwater,FL,33755,27.966235,-82.787412,5.0,10,General Dentistry; Dentists; Health & Medical; Cosmetic Dentists
0bPLkL0QhhPO5kt1_EXmNQ,Zio's Italian Market,2575 E Bay Dr,Largo,FL,33771,27.9161159,-82.7604608,4.5,100,Food; Delis; Italian; Bakeries; Restaurants
MUTTqe8uqyMdBl186RmNeA,Tuna Bar,205 Race St,Philadelphia,PA,19106,39.953949,-75.1432262,4.0,245,Sushi Bars; Restaurants; Japanese
rBmpy_Y1UbBx8ggHlyb7hA,Arizona Truck Outfitters,625 N Stone Ave,Tucson,AZ,85705,32.2298719,-110.9723419,4.5,10,Automotive; Auto Parts & Supplies; Auto Customization
M0XSSHqrASOnhgbWDJIpQA,Herb Import Co,712 Adams St,New Orleans,LA,70118,29.9414679565,-90.129952757,4.0,5,Vape Shops; Tobacco Shops; Personal Shopping; Vitamins & Supplements; Shopping
```

Data Sample

- Tips Data:

```
rca9751_nyu_edu@nyu-dataproc-m:~/project$ hadoop fs -cat project/output/part-r-00000 | head
4tF1CWdMxvvwpUIgGsDygA,_cb1Vg1NIWry8UA0jyuXnQ,Food is good value but a bit hot,2021-12-07 22:30:00,0
ckqKGM2hl7I9Chp5IpAhkw,s2eyoTuJrcP7I_XyjdhUHQ,Great pizza great price,2021-11-20 16:11:44,0
v48Spe6WEpqehsF2xQADpg,hYnMeAO77RGyTtIzUSKYzQ,Love their Cubans,2021-11-05 13:18:56,0
1uxtQAuJ2T5Xwa_wp7kUnA,OaGf0Dp56ARhQwIDT90w_g,Great food and service,2021-10-30 11:54:36,0
eYodOTF8pkqKPzHkcxZs-Q,3lHTewuKFt5IImbXJoFeDQ,Disappointed in one of your managers,2021-09-11 19:18:57,0
FowxkbAixI3hlREeCgIa_Q,kfNv-JZpuN6TVNSO6hHdkw,Great experience with a phenomenal food with a lot of flavor and a affordable price,2021-05-09 23:21:10,0
5hJR7ljJbhFgOaLi8iz5pQ,AXC__4yZrn-N3BT7-2bV_Q,BOMB food Super delicious great outdoor space and kid friendly,2021-05-04 21:44:53,0
2-vAo2UfkD7QHA5TG8kwmg,wQUBiBqlzC6cbdkX-GaBqQ,The food was delicious,2021-03-12 00:15:07,0
Apfz4xUeBOObgXkOhOnMxQ,90pJu2O7fIEm_N31Fyue7A,Great food cocktail ambience and service,2021-02-14 15:02:38,0
```

Data Flow


users
businesses
review
tips
HDFS
MapReduce
MapReduce
MapReduce
MapReduce
users-clean
businesses-clean
review-clean
tips-clean
HDFS
Hive
Trino
Analytics and Reports

Code Challenges

- Review Dataset: Having the review_id as the output key in MapReduce job made it hard to convert the text file to SQL dataset. Solution : Used a NullWritable class as the output key and ReviewDataWritableClass as the output value.

```
    @Override
    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
        try {
            // Parse JSON record
            JsonNode jsonNode = mapper.readTree(value.toString());

            // Extract fields from JSON record
            String reviewId = jsonNode.get("review_id").asText();
            String userId = jsonNode.get("user_id").asText();
            String businessId = jsonNode.get("business_id").asText();
            double stars = jsonNode.get("stars").asDouble();
            int useful = jsonNode.get("useful").asInt();
            int funny = jsonNode.get("funny").asInt();
            int cool = jsonNode.get("cool").asInt();
            String text = jsonNode.get("text").asText();
            String date = jsonNode.get("date").asText();
            ReviewDataWritable data = new ReviewDataWritable(new Text(reviewId), new Text(userId), new Text(b
usinessId), new DoubleWritable(stars), new DoubleWritable(useful), new DoubleWritable(funny), new DoubleWrita
ble(cool));
            context.write(NullWritable.get(), data);
        } catch (Exception e) {
            // Ignore invalid JSON records
        }
```

Code Challenges

- User Dataset - For the original json dataset, used a tuple class to store attributes in a clean and concise manner for map reduce jobs.

```
presto:aa9381_nyu_edu> SELECT users.userid, users.name, COUNT(DISTINCT business.city) AS num_locations_reviewed
                    -> FROM users
                    -> JOIN reviews ON users.userid = reviews.userid
                    -> JOIN business ON reviews.businessid = business.businessId
                    -> GROUP BY users.userid, users.name
                    -> ORDER BY num_locations_reviewed DESC
                    -> LIMIT 10;
                    ->
         userid          |   name   | num_locations_reviewed
-------------------------+----------+-----------------------
 _BcWyKQL16ndpBdggh2kNA | Karen    |                   173
 vmUqcqMjlWoBM6qfmUXgyQ | James    |                   147
 -G7Zkl1wIWBBmD0KRy_sCw | Gerald   |                   141
 XzpJ4uHkxARCFQiZ9bffyg | P        |                   121
 RCZ5M9o2-fxgFuurpmEs3w | Craig    |                   117
 FlXBpK_YZxLo27jcMdII1w | Mallory  |                   111
 GcdYgbaF75vj7RO6EZhPOQ | Kathleen |                   103
 pppIHoA8b8B8Wd5t72sDxA | Veronica |                   102
 6s-g2vFu12OemhiK3FJuOQ | Dave     |                   100
 lRRuTimITgwzoXLIM3g9qw | Tim      |                    99
(10 rows)

Query 20230502_175501_00239_d2d8x, FINISHED, 2 nodes
Splits: 359 total, 359 done (100.00%)
4.89 [9.13M rows, 644MB] [1.87M rows/s, 132MB/s]
```

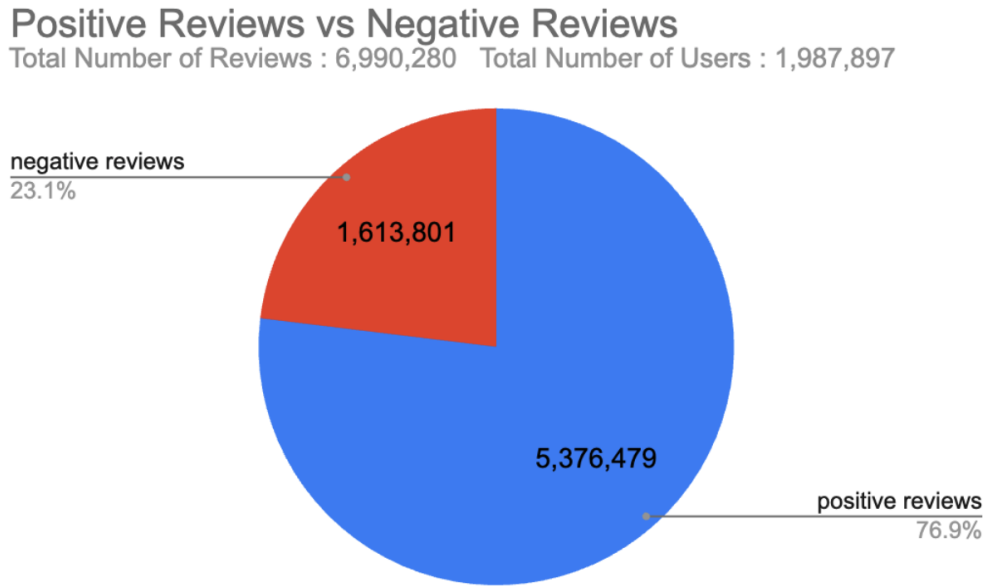
- Joining multiple tables, with similar fields and keeping track of their context.

Code Challenges

- Business dataset: Categories column is a string of comma-separated unordered categories; making it harder to query for category-based insights against other columns. Custom map reduce programs were used instead

business_id	catgeories
Pns2l4eNsfO8kk83dixA6A	Doctors; Traditional Chinese Medicine; Naturopathic/Holistic; Acupuncture; Health & Medical; Nutritionists
mpf3x-BjTdTEA3yCZrAYPw	Shipping Centers; Local Services; Notaries; Mailbox Centers; Printing Services
tUFrWirKiKi_TAnsVWINQQ	Department Stores; Shopping; Fashion; Home & Garden; Electronics; Furniture Stores
MTSW4McQd7CbVtyjqoe9mw	Restaurants; Food; Bubble Tea; Coffee & Tea; Bakeries
mWMc6_wTdE0EUBKIGXDVfA	Brewpubs; Breweries; Food
CF33F8-E6oudUQ46HnavjQ	Burgers; Fast Food; Sandwiches; Food; Ice Cream & Frozen Yogurt; Restaurants
n_0UpQx1hsNbnPUSlodU8w	Sporting Goods; Fashion; Shoe Stores; Shopping; Sports Wear; Accessories
qkRM_2X51Yqxk3btlwAQIg	Synagogues; Religious Organizations
k0hlBqXX-Bt0vf1op7Jr1w	Pubs; Restaurants; Italian; Bars; American (Traditional); Nightlife; Greek
bBDDEgkFA1Otx9Lfe7BZUQ	Ice Cream & Frozen Yogurt; Fast Food; Burgers; Restaurants; Food
UJsufbvfyfONHeWdvAHKjA	Department Stores; Shopping; Fashion
eEOYSgkmpB90uNA7lDOMRA	Vietnamese; Food; Restaurants; Food Trucks
il_Ro8jwPlHresjw9EGmBg	American (Traditional); Restaurants; Diners; Breakfast & Brunch
jaxMSoInw8Poo3XeMJt8lQ	General Dentistry; Dentists; Health & Medical; Cosmetic Dentists
0bPLkL0QhhPO5kt1_EXmNQ	Food; Delis; Italian; Bakeries; Restaurants

Results and Insights

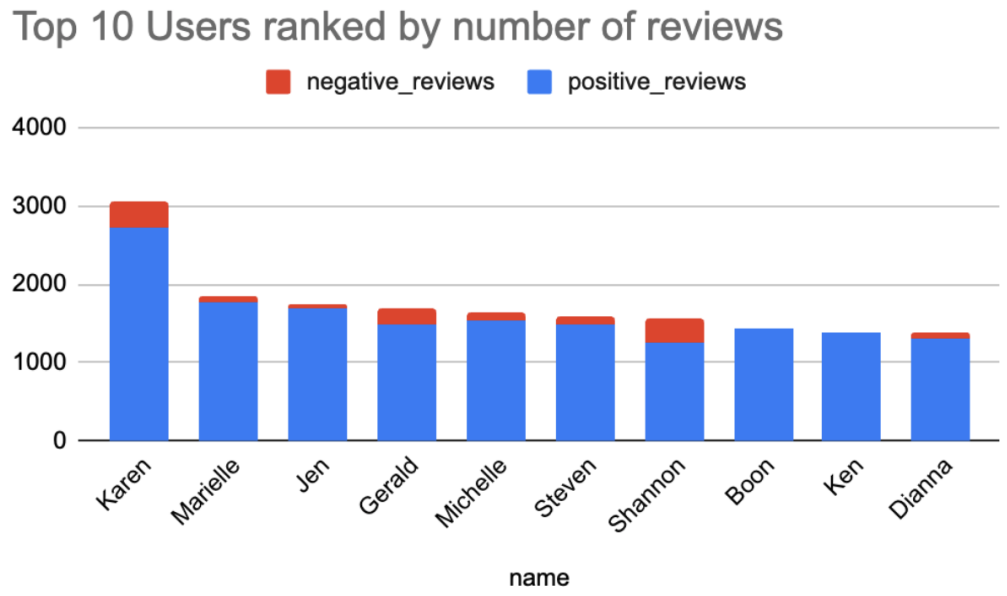


Total Number of Users : 1,987,897
Total Number of Reviews : 6,990,280
Average number of reviews written by a single user : 3.51
Average stars given : 3.748

Top 10 Users ranked by Number of Reviews Written

user_id	name	average_stars	total_reviews	positive_reviews	negative_reviews
_BcWyKQL16ndpBdggh2kNA	Karen	3.6377952	3048	2715	333
Xw7ZjaGfr0WNVt6s_5KZfA	Marielle	4.072826	1840	1765	75
0lgx-a1wAstiBDerGxXk2A	Jen	3.990269	1747	1705	42
-G7Zkl1wIWBBmD0KRy_sCw	Gerald	3.6527944	1682	1499	183
ET8n-r7glWYqZhuR6GcdNw	Michelle	4.0465817	1653	1537	116
bYENop4BuQepBjM1-Bl3fA	Steven	3.8536122	1578	1493	85
1HM81n6n4iPIFU5d2Lokhw	Shannon	3.0450451	1554	1257	297
fr1Hz2acAb3OaL3l6DyKNg	Boon	3.9467864	1447	1425	22
wXdbkFZsfDR7utJvbWElyA	Ken	4.210602	1396	1387	9
Um5bfs5DH6eizgjH3xZsvg	Dianna	3.8044572	1391	1298	93

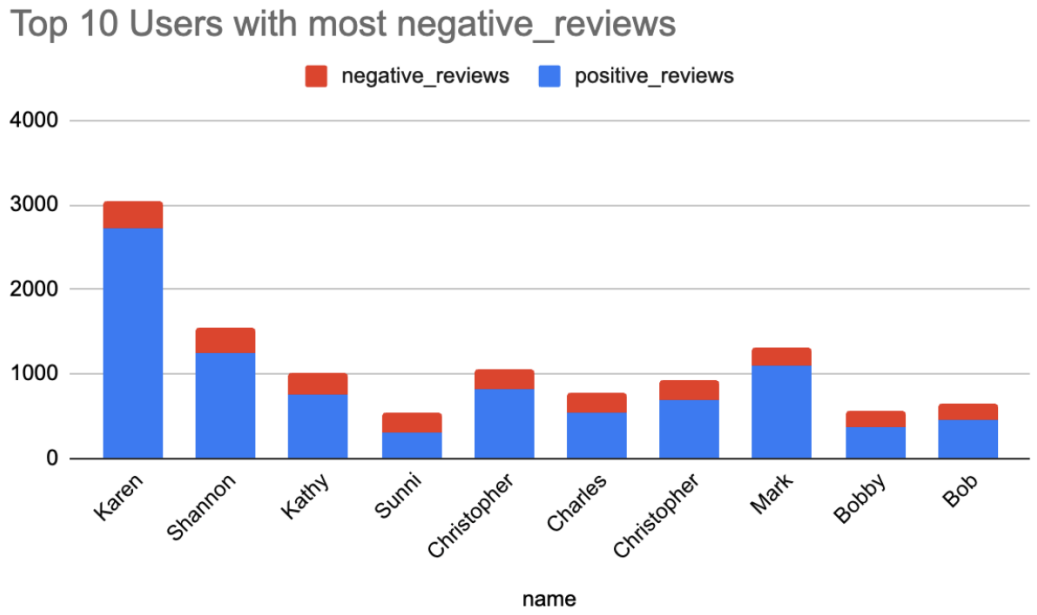
Results and Insights Cont.

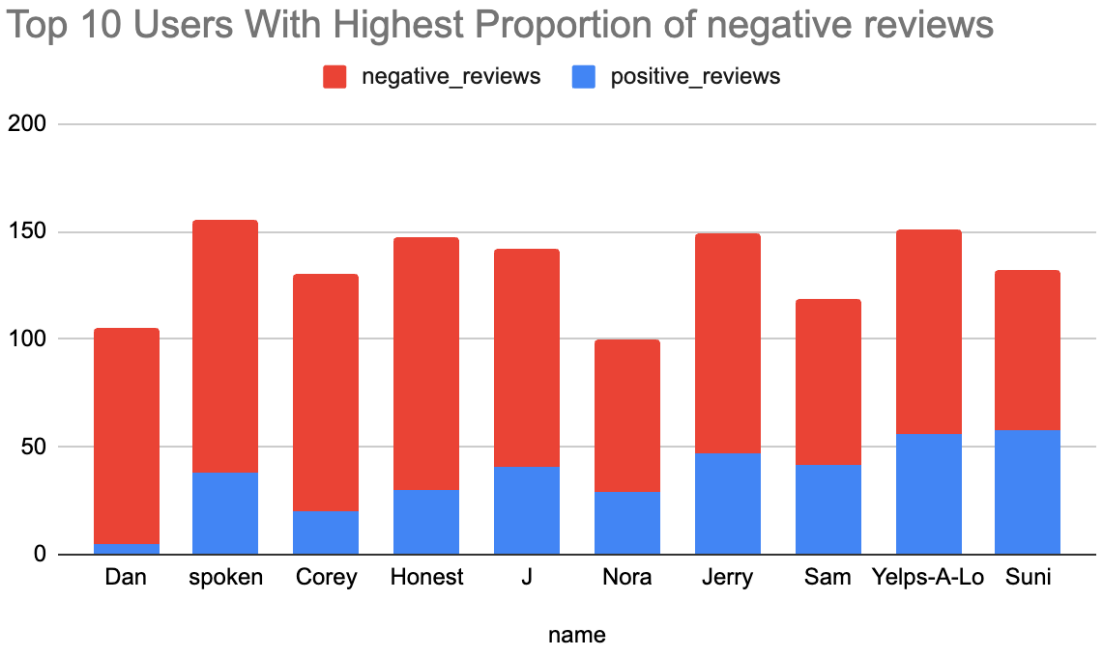


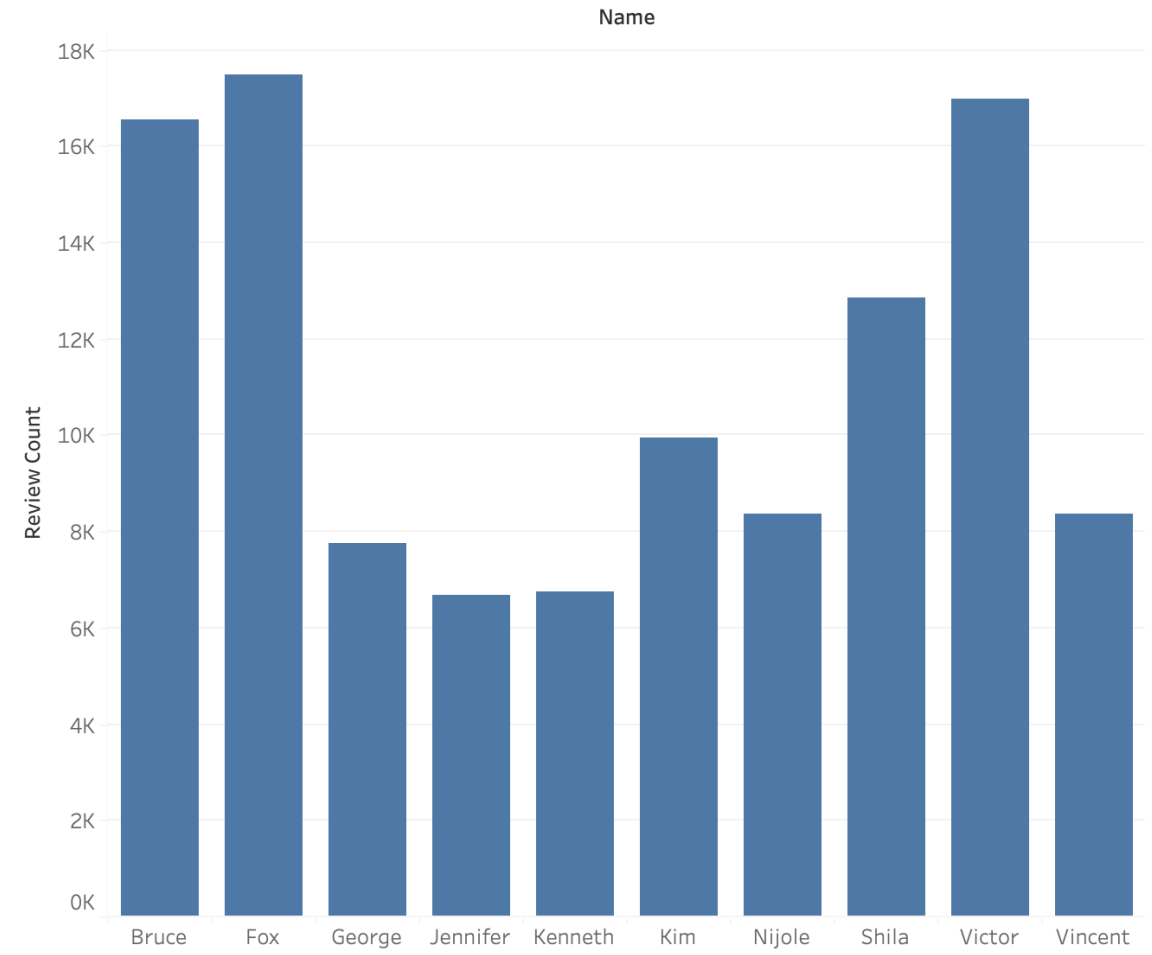
Top 10 Users ranked by Number of Reviews Written

user_id	name	average_stars	total_reviews	positive_reviews	negative_reviews
_BcWyKQL16ndpBdggh2kNA	Karen	3.6377952	3048	2715	333
Xw7ZjaGfr0WNVt6s_5KZfA	Marielle	4.072826	1840	1765	75
0lgx-a1wAstiBDerGxXk2A	Jen	3.990269	1747	1705	42
-G7Zkl1wlWBBmD0KRy_sCw	Gerald	3.6527944	1682	1499	183
ET8n-r7glWYqZhuR6GcdNw	Michelle	4.0465817	1653	1537	116
bYENop4BuQepBjM1-Bl3fA	Steven	3.8536122	1578	1493	85
1HM81n6n4iPlFU5d2Lokhw	Shannon	3.0450451	1554	1257	297
fr1Hz2acAb3OaL3l6DyKNg	Boon	3.9467864	1447	1425	22
wXdbkFZsfDR7utJvbWElyA	Ken	4.210602	1396	1387	9
Um5bfs5DH6eizgjH3xZsvg	Dianna	3.8044572	1391	1298	93

Results and Insights Cont.





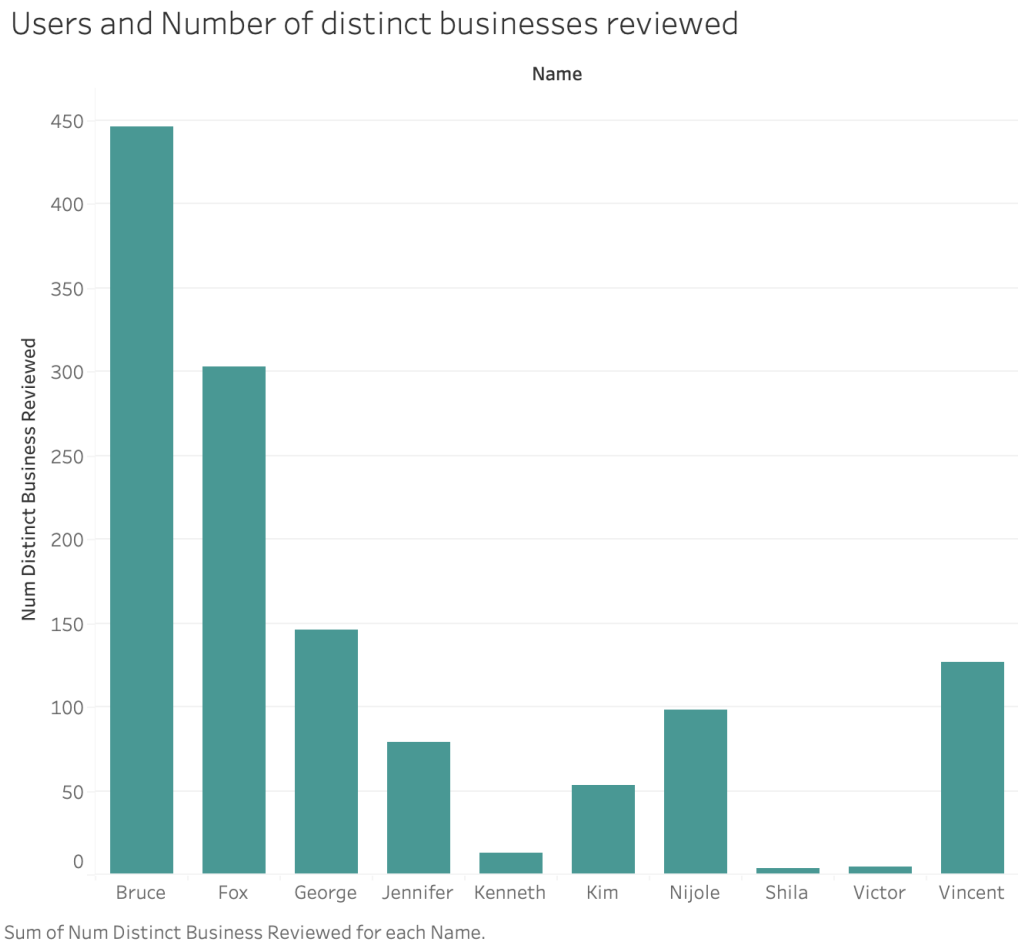


Sum of Review Count for each Name.

Top ten users by Review Count

userId	name	review_count
Hi10sGSZNxQH3NLyWSZ1oA	Fox	17473.0
8k3aO-mPeyhbR5HUucA5aA	Victor	16978.0
hWDybu_KvYLSdEFzGrniTw	Bruce	16567.0
RtGqdDBvvBCjcu5dUqwfzA	Shila	12868.0
P5bUL3Engv-2z6kKohB6qQ	Kim	9941.0
nmdkHL2JKFx55T3nq5VziA	Nijole	8363.0
bQCHF5rn5lMI9c5kEwCaNA	Vincent	8354.0
8RcEwGrFIgkt9WQ35E6SnQ	George	7738.0
Xwnf20FKuikiHcSpcEbpKQ	Kenneth	6766.0
CxDOIDnH8gp9KXzpBHJYXw	Jennifer	6679.0

(10 rows)



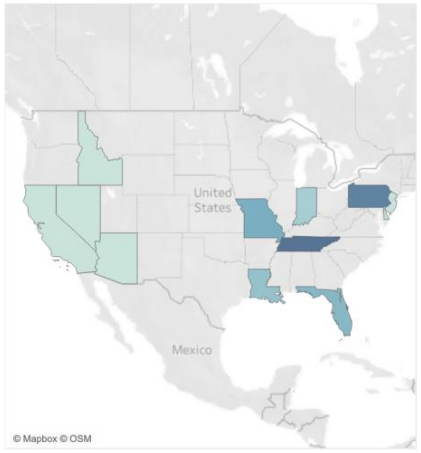
Top ten users and the number of number of distinct businesses reviewed by them.

userid	name	review_count	num_distinct_businesses_reviewed
Hi10sGSZNxQH3NLyWSZ1oA	Fox	17473.0	303
8k3aO-mPeyhbR5HUucA5aA	Victor	16978.0	5
hWDybu_KvYLSdEFzGrniTw	Bruce	16567.0	446
RtGqdDBvvBCjcu5dUqwfzA	Shila	12868.0	4
P5bUL3Engv-2z6kKohB6qQ	Kim	9941.0	53
nmdkHL2JKFx55T3nq5VziA	Nijole	8363.0	98
bQCHF5rn5lMI9c5kEwCaNA	Vincent	8354.0	127
8RcEwGrFIgkt9WQ35E6SnQ	George	7738.0	146
Xwnf20FKuikiHcSpcEbpKQ	Kenneth	6766.0	13
CxDOIDnH8gp9KXzpBHJYXw	Jennifer	6679.0	79

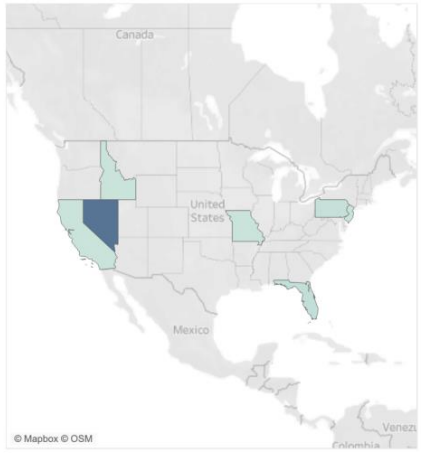
(10 rows)

Yelp Review Analysis (Group 12)

1. Fox - Review Count 17473 , Number of Distinct Businesses Reviewed 303



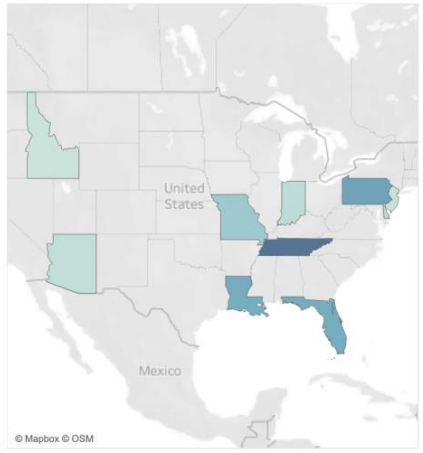
5. Kim - Review Count 9941 , Number of Distinct Businesses Reviewed 53



10. Jennifer - Review Count 6679 , Number of Distinct Businesses Reviewed 79

4. Shila - Review Count 12868 , Number of Distinct Businesses Reviewed 4

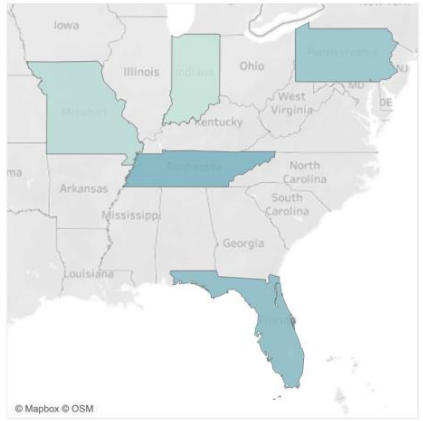
3. Bruce - Review Count 16567 , Number of Distinct Businesses Reviewed 446



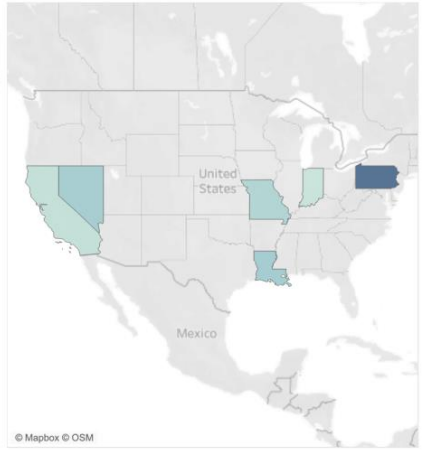
2. Victor - Review Count 16978 , Number of Distinct Businesses Reviewed 5

6. Nijole - Review Count 8363 , Number of Distinct Businesses Reviewed 98

9. Kenneth - Review Count 6766 , Number of Distinct Businesses Reviewed 13



7. Vincent - Review Count 8354 , Number of Distinct Businesses Reviewed 127



8. George - Review Count 7738 , Number of Distinct Businesses Reviewed 146

Analysed the top ten users who wrote most reviews, queried the distinct businesses they wrote reviews for. Lastly, displaying the cities (location) these businesses belong to.

Results and Insights Cont.

category	count
restaurants	52268
food	27781
shopping	24395
home services	14356
beauty & spas	14292
nightlife	12281
health & medical	11890
local services	11198
bars	11065
automotive	10773
event planning & services	9895
sandwiches	8366
american (traditional)	8139
active life	7687
pizza	7093
coffee & tea	6703
fast food	6472
breakfast & brunch	6239
american (new)	6097
hotels & travel	5857

Top 20 categories by Review Count

city	avg_stars	num_reviews
Philadelphia	3.6230352	936240
New Orleans	3.8226767	621361
Nashville	3.6377852	441053
Tampa	3.583315	439506
Tucson	3.594919	387254
Indianapolis	3.5797083	349228
Reno	3.7615838	334610
Santa Barbara	4.0514493	262853
Saint Louis	3.5940542	244360
Boise	3.714164	101893
Edmonton	3.439058	98204
Clearwater	3.6013057	84190
Saint Petersburg	3.7143717	76219
Sparks	3.6474755	69567
Metairie	3.493305	61970
St. Louis	3.6653388	61270
Franklin	3.6053748	54785
St. Petersburg	3.8	52620
Goleta	3.7493734	44126
Wilmington	3.423928	43005

Average stars of 20 most popular cities

Results and Insights Cont.

state	avg_stars	num_reviews
PA	3.5730193	1540790
FL	3.6109571	1119926
LA	3.6791615	743176
TN	3.5714996	598195
MO	3.5460918	483897
IN	3.5882459	472565
AZ	3.5920098	412639
NV	3.7368762	409950
CA	3.9967327	339637
NJ	3.4591143	249837

Average stars of top 10 states

name	city	review_count
Acme Oyster House	New Orleans	7568
Hattie B's Hot Chicken - Nashville	Nashville	6093
Reading Terminal Market	Philadelphia	5721
Pappy's Smokehouse	Saint Louis	3999
Los Agaves	Santa Barbara	3834
Grand Sierra Resort and Casino	Reno	3345
Datz	Tampa	3260
Frenchy's Rockaway Grill	Clearwater Beach	2301
The Eagle	Indianapolis	2233
Prep & Pastry	Tucson	2126

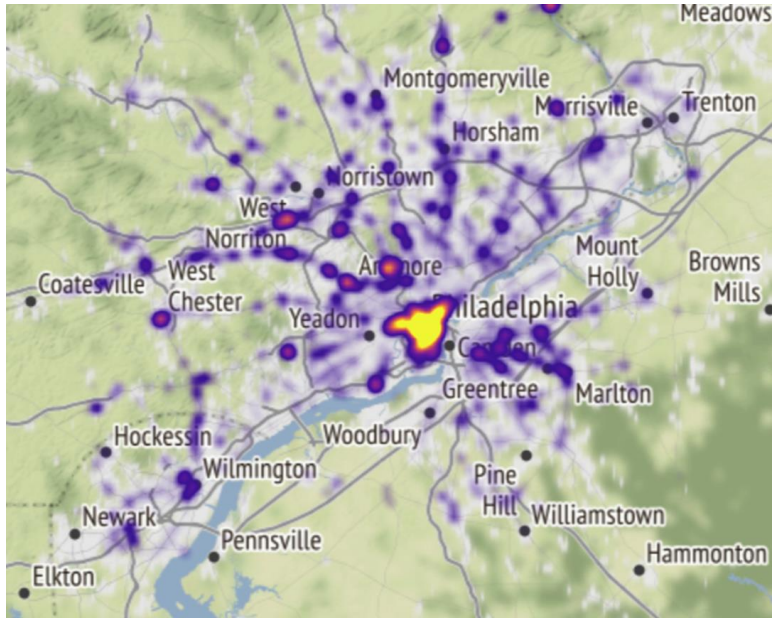
Most popular businesses in top 10 cities by review count

Results and Insights Cont.

North American heatmap
based on review count

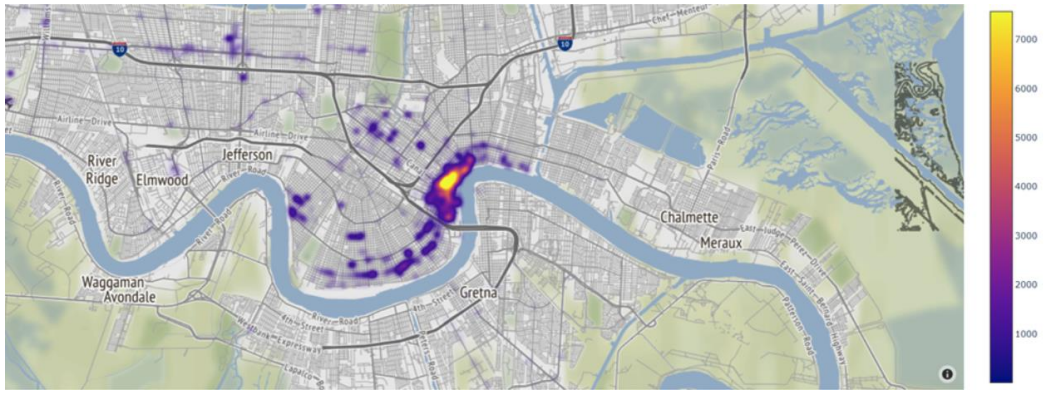
Results and Insights Cont.

Review Count heatmap for top 3 most popular cities

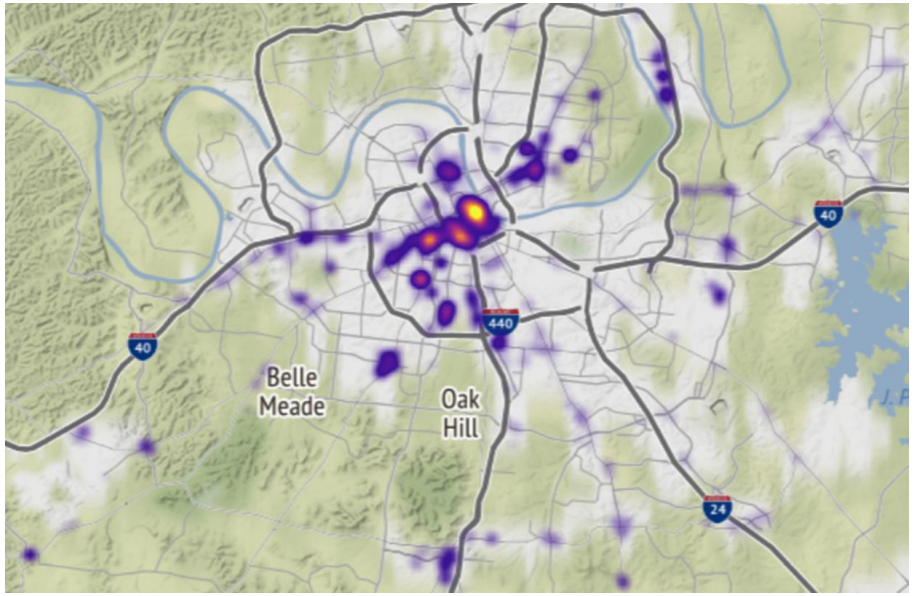


Philadelphia, PA

New Orleans, LA



Nashville, TN



Obstacles

- Reduce-side join was too slow as the datasets were large. Therefore, we used Hive and Trino to join the tables
- Having four datasets, with similar attributes, data schema became a bit confusing since few tables had same column names with different meaning. For example, review_count field for business and user datasets.
- Wrote a few queries which didn't result in any valuable insights, since the data became very sparse.

Summary

- From the analysis of the users, review and business dataset, it is easier to find the locations most active users are writing reviews from, since user dataset does not have any location attribute. This can be used drive targeted advertisements or recommendations to that particular user.
- We were able to find the top users who wrote the most reviews and the most negative reviews. There were some users whose reviews were 95% negative. Yelp could use this analysis to warn/block these users, as the star rating can affect business's reputation.
- Business analysis and heatmap shows us the most popular areas in the city, which can be used to decide optimal locations for a new business. Average city stars can be used as a benchmark to gauge business sentiment

Acknowledgments

- NYU HPC
- Tableau student license

References

- Kaggle Yelp Academic Dataset (link)
- Hadoop: The Definitive Guide, 4th Edition (link)

Thank You.