Yelp Dataset Analysis Report

CS686 Data Processing in Cloud - Personal Project

Zini Zhu

Motivation

In the past few years, Yelp's already become a must-have app for everyone. For me, I use it to search for good restaurants every week, and every time I can find something new. It's like a treasure land. With this personal project opportunity, the original idea was just to focus on restaurants and conduct a general analysis on the variety and popularity of different restaurant categories. But later when looking into Yelp's dataset, I found it has far more than that. I used to think that Yelp is just for restaurant recommendations, but it also covers other businesses. The dataset includes user information and reviews as well, and the addition of these data allows me to do more complex analysis. Finally I decided to focus on three areas: Yelp itself, businesses that advertise on Yelp, and Yelp's users. Hopefully with this data, I could take a closer look at Yelp's operation condition, and probably give out some useful advice on its future plan(WOW). By analyzing the businesses, users and reviews data, I could have a better understanding of their relationships, and see if it is really worth advertising on Yelp.

Data

The dataset is from Yelp's official website and four json files were downloaded from Kaggle (business/review/user/tip). Data is well formatted and clean, thus not requiring much sanity check, and here's examples for each json object:

- business(138MB)


```
{
   "business_id":"QXAEGFB4oINsVuTFxEYKFQ",
   "name":"Emerald Chinese Restaurant",
   "address":"30 Eglinton Avenue W",
   "city":"Mississauga",
   "state":"ON",
   "postal_code":"L5R 3E7",
   "latitude":43.6054989743,
   "longitude":-79.652288909,
   "stars":2.5,
   "review_count":128,
   "is_open":1,
   "attributes":{
      "RestaurantsReservations":"True",
      "GoodForMeal":"{'dessert': False, 'latenight': False, 'lunch': True, 'dinner': True, 'brunch': False, 'breakfast': False}",
      "BusinessParking":"{'garage': False, 'street': False, 'validated': False, 'lot': True, 'valet': False}",
      "Caters":"True",
      "NoiseLevel":"u'loud'",
      "RestaurantsTableService":"True",
      "RestaurantsTakeOut":"True",
      "RestaurantsPriceRange2":"2",
```

```
      "OutdoorSeating":"False",
      "BikeParking":"False",
      "Ambience":"{'romantic': False, 'intimate': False, 'classy': False, 'hipster': False, 'divey': False, 'touristy': False,
'trendy': False, 'upscale': False, 'casual': True}",
      "HasTV":"False",
      "WiFi":"u'no'",
      "GoodForKids":"True",
      "Alcohol":"u'full_bar'",
      "RestaurantsAttire":"u'casual'",
      "RestaurantsGoodForGroups":"True",
      "RestaurantsDelivery":"False"
   },
   "categories":"Specialty Food, Restaurants, Dim Sum, Imported Food, Food, Chinese, Ethnic Food, Seafood",
   "hours":{
      "Monday":"9:0-0:0",
      "Tuesday":"9:0-0:0",
      "Wednesday":"9:0-0:0",
      "Thursday":"9:0-0:0",
      "Friday":"9:0-1:0",
      "Saturday":"9:0-1:0",
      "Sunday":"9:0-0:0"
   }
}
```


Business.json contains business id, name, geo info, stars, categories, open hours and other attributes.

- tips(244MB)


```
{
   "user_id":"UPw5DWs_b-e2JRBS-t37Ag",
   "business_id":"VaKXUpmWTTWDKbpJ3aQdMw",
   "text":"Great for watching games, ufc, and whatever else tickles yer fancy",
   "date":"2014-03-27 03:51:24",
   "compliment_count":0
}
```


Tips.json contains a tip as text for one business, as well as the compliment count.

- review(5.35GB)


```
{
   "review_id":"Q1sbwvVQXV2734tPgoKj4Q",
   "user_id":"hG7b0MtEbXx5QzbzE6C_VA",
   "business_id":"ujmEBvifdJM6h6RLv4wQIg",
   "stars":1.0,
   "useful":6,
   "funny":1,
   "cool":0,
   "text":"Total bill for this horrible service? Over $8Gs. These crooks actually had the nerve to charge us $69 for 3
pills. I checked online the pills can be had for 19 cents EACH! Avoid Hospital ERs at all costs.",
   "date":"2013-05-07 04:34:36"
}
```


Reviews.json is the largest file among all files, it contains review content, reviewer, business and feedback.

- user(2.49GB)


```
{
   "user_id":"l6BmjZMeQD3rDxWUbiAiow",
   "name":"Rashmi",
   "review_count":95,
   "yelping_since":"2013-10-08 23:11:33",
```

```
    "useful":84,
    "funny":17,
    "cool":25,
    "elite":"2015,2016,2017",
    "friends":"c78V-rj8NQcQjOI8KP3UEA, alRMgPcngYSCJ5naFRBz5g",
    "fans":5,
    "average_stars":4.03,
    "compliment_hot":2,
    "compliment_more":0,
    "compliment_profile":0,
    "compliment_cute":0,
    "compliment_list":0,
    "compliment_note":1,
    "compliment_plain":1,
    "compliment_cool":1,
    "compliment_funny":1,
    "compliment_writer":2,
    "compliment_photos":0
}
```


User.json records each user's name, review count, fans, and feedback on his/her reviews.

Data Pre-processing

I directly downloaded the json files from Kaggle. I wrote a Python script to extract needed fields, trim white space, convert the date field to timestamp and reformat the json so that I could upload json files directly to BigQuery. It's worth mentioning that I could've better formalize the string fields, for example, converting strings to lowercase, so that I wouldn't have to handle it later in all queries.

Python function to extract fields:

```python
import json
from datetime import datetime


def json_parse_tip(input, output):
  fp = open(output, 'w')
  for line in open(input, mode='r'):
      obj = json.loads(line)
      date_str = obj['date']
      date_time = datetime.fromisoformat(date_str)
      timestamp = (int)(datetime.timestamp(date_time))
      obj['date'] = timestamp
      del obj['compliment_count']
      json.dump(obj, fp)
      fp.write('\n')
json_parse_business("./test.json", './res.json')
```

BigQuery

a. How to load data to BigQuery?

I uploaded the json files to GCS, and then loaded them to BigQuery.

Gsutil command to upload files:

```
$gsutil cp /Users/zhuzini/reviews.json gs://yelp-food-hunter-dataset
```

BQ command to create new dataset:

```
$bq make yelp-data-set
```

BQ command to upload tables:

```
$bq load yelp-data-set.tips tips.json user_id:string,business_id:string,text:string,date:timestamp
```

b. Tables and schemas

I had 4 tables for business, reviews, tips and users, and here are what the schemas look like:

tips

Schema | Details | Preview

Field name	Type	Mode
user_id	STRING	NULLABLE
business_id	STRING	NULLABLE
text	STRING	NULLABLE
date	TIMESTAMP	NULLABLE

reviews

Schema | Details | Preview

Field name	Type	Mode
review_id	STRING	NULLABLE
user_id	STRING	NULLABLE
business_id	STRING	NULLABLE
stars	NUMERIC	NULLABLE
useful	INTEGER	NULLABLE
text	STRING	NULLABLE
date	TIMESTAMP	NULLABLE

business

Schema | Details | Preview

Field name	Type	Mode
business_id	STRING	NULLABLE
name	STRING	NULLABLE
address	STRING	NULLABLE
city	STRING	NULLABLE
state	STRING	NULLABLE
postal_code	STRING	NULLABLE
latitude	FLOAT	NULLABLE
longitude	FLOAT	NULLABLE
stars	NUMERIC	NULLABLE
review_count	INTEGER	NULLABLE
categories	STRING	REPEATED
hours	RECORD	NULLABLE
hours. **Monday**	STRING	NULLABLE
hours. **Tuesday**	STRING	NULLABLE
hours. **Wednesday**	STRING	NULLABLE
hours. **Thursday**	STRING	NULLABLE
hours. **Friday**	STRING	NULLABLE
hours. **Saturday**	STRING	NULLABLE
hours. **Sunday**	STRING	NULLABLE

users

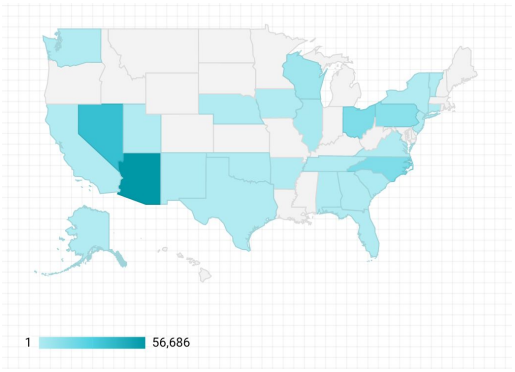
Schema | Details | Preview

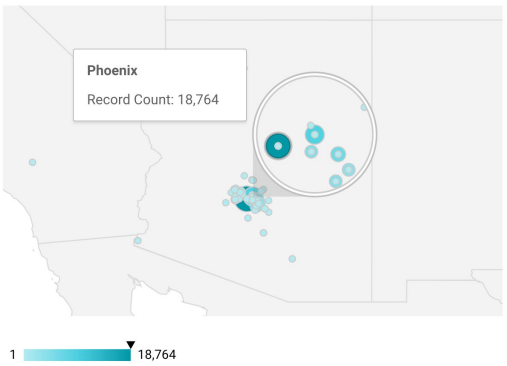
Field name	Type	Mode
user_id	STRING	NULLABLE
name	STRING	NULLABLE
review_count	INTEGER	NULLABLE
yelping_since	TIMESTAMP	NULLABLE
useful	INTEGER	NULLABLE
fans	INTEGER	NULLABLE
average_stars	NUMERIC	NULLABLE

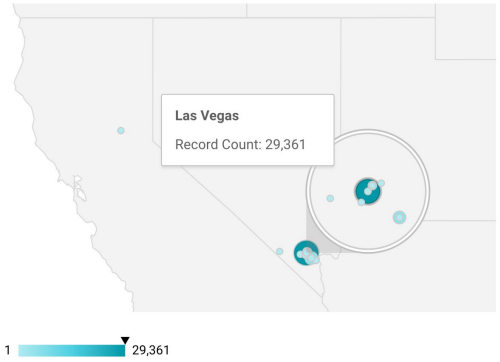
c. Visualization of data

- Total number of rows for each table:

Business#	Users#	Reviews#	Tips#
192,609	1,637,138	6,685,900	1,223,094



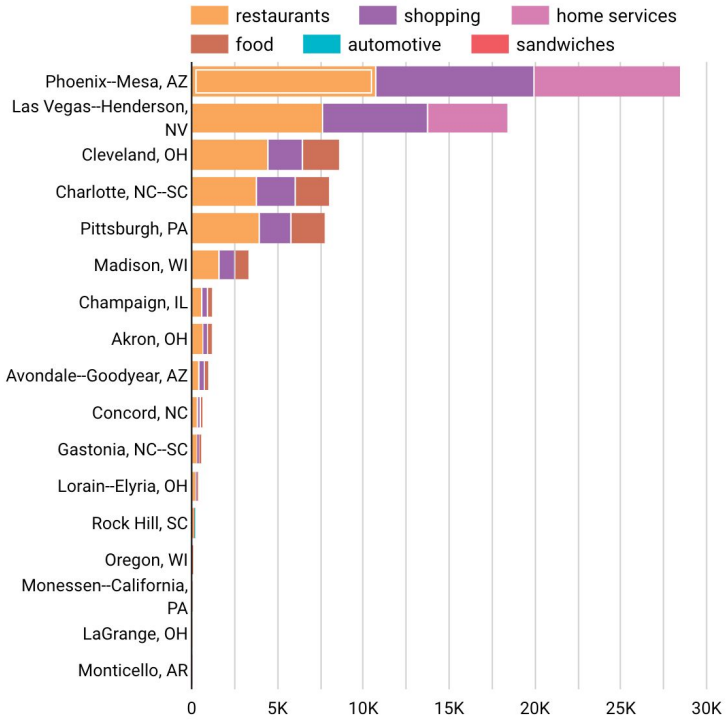




- Business Distribution

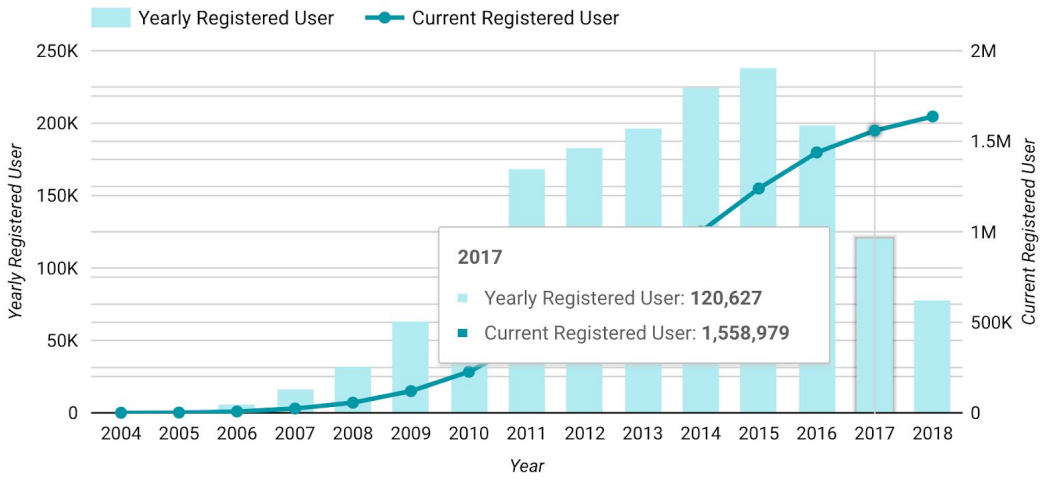
  The map on the left shows the geographical distribution of businesses included in the dataset.This dataset mainly includes the business information in the southern part of America, most gathering in Arizona and Nevada. If we drill down to Arizona, we can further discover that most business is in Phoenix. Same thing happens in Las Vegas, Nevada. I was planning to analyze the geological preference of businesses advertising on Yelp, but due to the regional concentration of this dataset, this analysis is not reasonable here.

  With the information in only a few cities, I studied the portion of 3 most popular business categories in these cities, and the result is in the below bar chart.

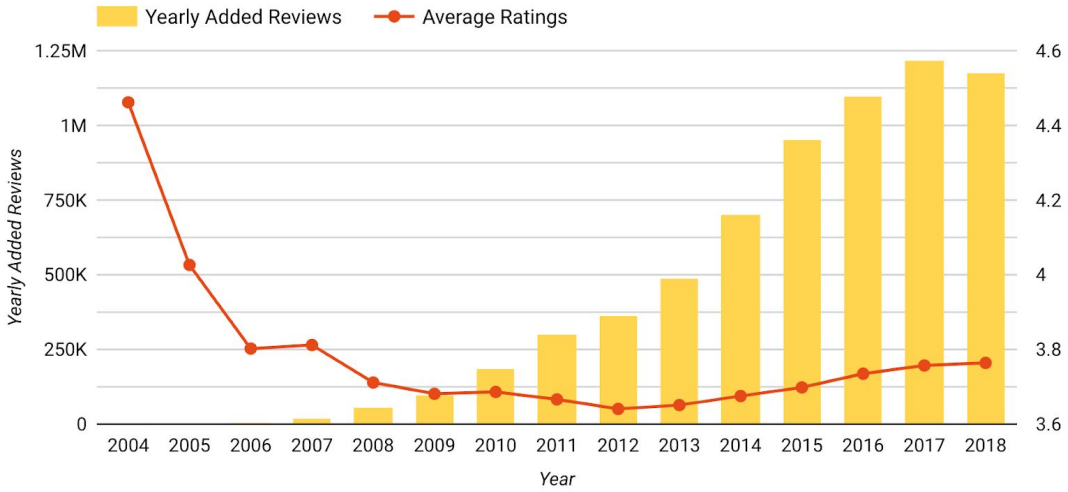


"Restaurants" or food-related categories are the most popular category in all 17 cities. Besides, in larger cities, different businesses tend to distribute evenly, while in smaller cities, food-related businesses take a fairly big portion.

- Yearly new-registered user



This chart plots the total number of registered users & yearly new-registered users in each year. As we can see in 2015 the increase in new-registered users reaches its peak. Yelp has definitely succeeded in the past 15 years, but due to the fall after 2015, it needs to work harder to attract more potential users.



- Reviews

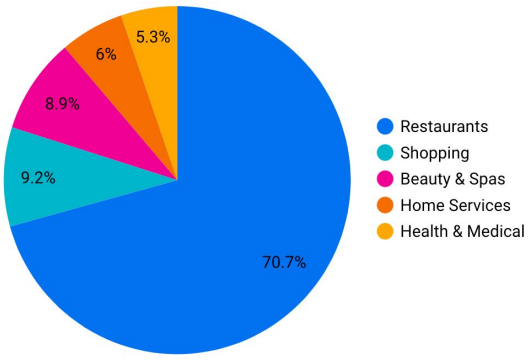
  The bars in this chart reveal the number of reviews made every year. Unlike the number of new users, the number of new reviews keeps increasing, and exceeds 1M after 2015. From this we can see an increase in user activity, but even at 2017, the average reviews each user makes in one year(based on this dataset) is just ~1, and Yelp should figure out better strategies to increase this number.

d. Analysis

- Yelp

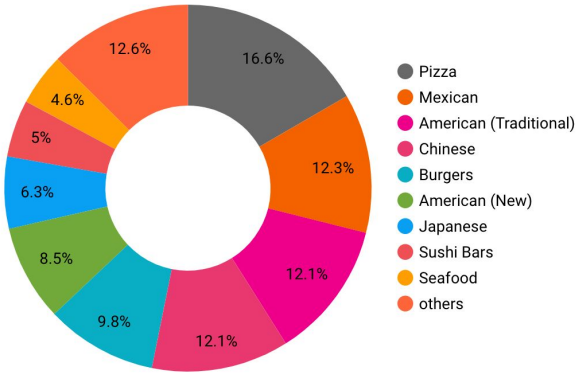
  In section **c** we've already made some analysis on Yelp's operation condition in the past 15 years. Here I further categorize the types of businesses based on the categories field in the business table. There are 1300 different categories in total (each business can have multiple

tags), for convenience I trimmed those detailed tags and just counted 5 main categories, representing 5 main business fields.



As we can see from the pie chart, food businesses account for ~70% of all businesses. Combined with the results from the above bar chart, it is obvious that restaurants prefer to advertise on Yelp in all cities, and the categories are more varied in larger cities. With a large user base, Yelp can think of how to attract businesses from other fields.

Just for fun, let's take a look at the variety of restaurants Yelp has:



- Business

  For analysis towards business, I mainly studied on these topics:
  - Is it useful for businesses to advertise on Yelp?
  - What's users' feedback on each business?
  - What's the reliability of these reviews and ratings?

  For generosity, I picked 3 popular categories, and 2 businesses with the most review count for each category, and conducted the above analysis on the selected restaurants. A summary report is generated for each restaurant using Google Data Studio.

Business Name	Category	Review Count
Bacchanal Buffet	Restaurants	8339
Mon Ami Gabi	Restaurants	8348
The Buffet	Beauty & Spas	4400
Trump International Hotel...	Beauty & Spas	1842
Gold & Silver Pawn Shop	Shopping	799
Rio All Suites Hotel & Casi...	Shopping	2498

Let’s pick one report and take a closer look at it:

Mon Ami Gabi

categories

Food, French, Breakfast & Brunch

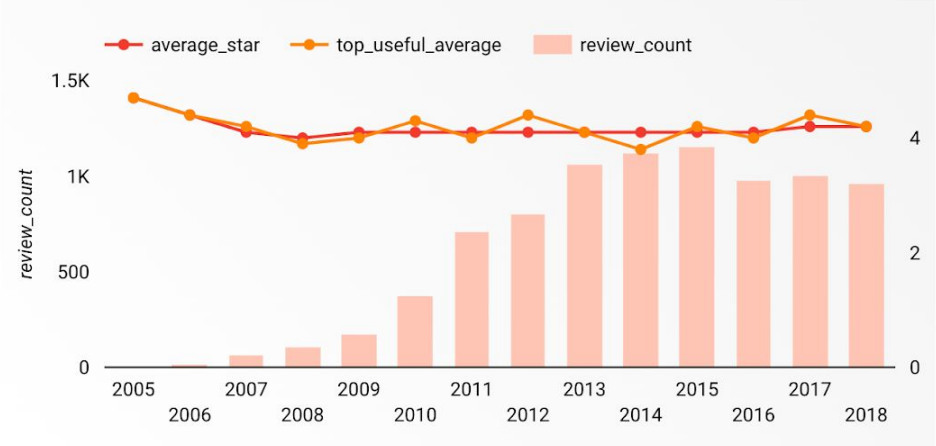
Summary

Useful Reviews
2,595

Average Stars
4

On-Yelp-Years
13

Stars / Reviews Number Trend



Review Reliability Stats

On-Yelp Years	Reviews #	Review Users #	Useful per Review	Star Std Deviation	Suspicious User
13	8,570	8,349	0.8	1	1

	Review	stars	useful	month	year
1.	One of my favorite restaurants to visit on the Strip whenever I am in Vegas. It's busy, so make reservations, but it's busy for good reason. Excellent steaks, seafood, and desserts, wonderful service and a great ambiance even on busy nights. I highly recommend the Beef Tournado medallions with truffle butter and your choice of sauce (I personally prefer the blue cheese Roquefort, but any of the others would be just as tasty). The prime steak is a flattened but large piece of meat, plenty for one person. Their filet mignon is a nice 8 oz portion. Definitely get their mushrooms of the side dish options, and get seconds on their bread :) Their profiteroles are huge, so make sure you share those if you order them.	5	2	November	2018
2.	I really love the Las Vegas Strip and I have eaten at Mon Ami Gabi so many times! This is really an awesome restaurant at many levels! I love the atmosphere in the morning for a nice breakfast watching the Bellagio famous fountains directly across Las Vegas Blvd with a great view from the patio part of the restaurant! Talk about Ambience. The atmosphere sitting on the patio is awesome enjoying the morning sun and on the inside very beautiful for fine dining! The staff is very professional, the restaurant is very clean, the tables are set of high class, the prices reasonable for being on the Las Vegas Strip and totally a 5 star service from the servers. The food is delicious "YUM" and food comes to the table with great	5	55	October	2018

The report contains four main parts:

- A summary of typical stats
- Rating and total review number’s’ changes along the years
- Calculated stats to evaluate the reliability of the reviews
- Top reviews

And let’s go over each part.

- Summary of typical stats
  In this section the report displays three statistics: total number of useful reviews, years that this business has been advertising on Yelp, and the average stars it earns during the years. Picking reviews with a “useful” tag instead of counting all reviews can give users a better sense of how helpful these reviews can be. A combination of average stars and business’s years on Yelp is a good representation of business reputation. For example in the above report, this business keeps a good rating of 4 over the past 12 years, so we can safely consider it as a good business.

- Rating and total review number’s changes along the years
  The bars in the chart represent the number of reviews made each year, and all businesses that I studied show a similar trend: it increases in the first few years and then maintains stable in the following years. This shows that advertising on Yelp helps increase the business popularity at the beginning.

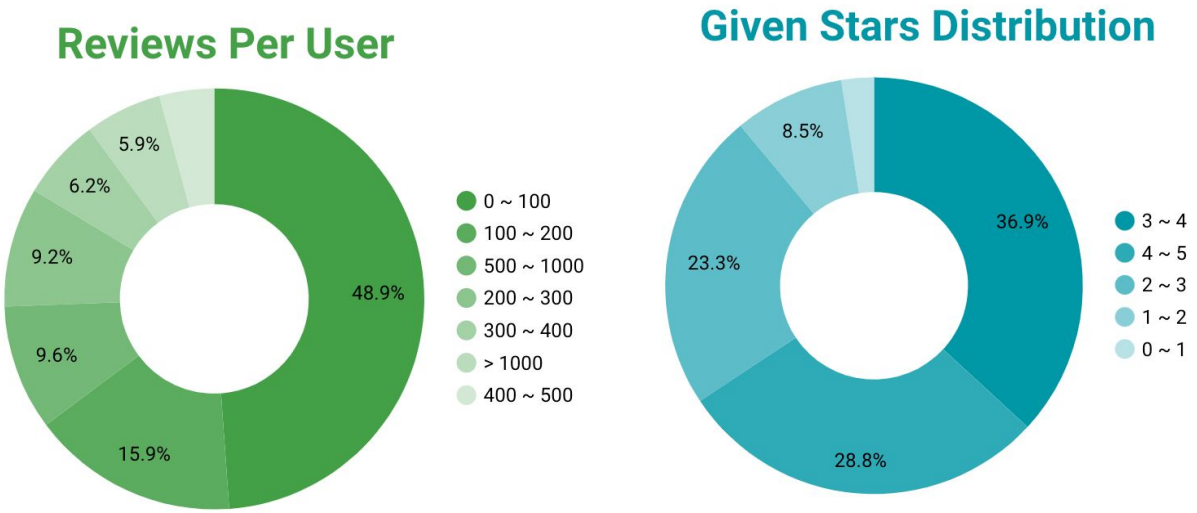
  The two lines in the chart represent both the average stars calculated from all reviews, and the average stars calculated from the top 20 reviews with the most ‘useful’ click. I did this to see if the most useful reviews can represent the opinions of all users. The results(for th 6 picked businesses) show that the two values are highly consistent. This means for users, it’s useful to look at the top reviews, and for businesses, it’s important for them to process the advice and complaints mentioned in the useful reviews well.

- Review Reliability Stats
  This table contains a few calculated fields that I used to evaluate whether the information for this business is worth reading.
    - On-Yelp-Years: again, if a business can survive on Yelp for long, then most likely it won’t be too bad.
    - Review and reviewers count: these show the popularity of the restaurant.
    - Useful per review: this is the average ‘useful click’ for each review on this business. The higher the value is, the more reliable these reviews are.
    - Star Standard Deviation: this is the standard deviation on stars, showing the variation among users’ ratings.
    - Suspicious user count: sometimes businesses may hire people or create bots to leave good comments and give high stars to improve its reputation, or its competitors may do so to deliberately leave bad comments. so I tried to identify those suspicious users based on this criterion: if a user gives more than 10 reviews with the same star for one business in a single year, then he/she is considered suspicious. For example in the above business one suspicious user is detected. But for a clearer comparison, I could’ve calculated the average of these stats for all businesses in the dataset that can be served as a reference. Also, the result can be more intuitive if I can create a “review quality score” for each business based on these data.

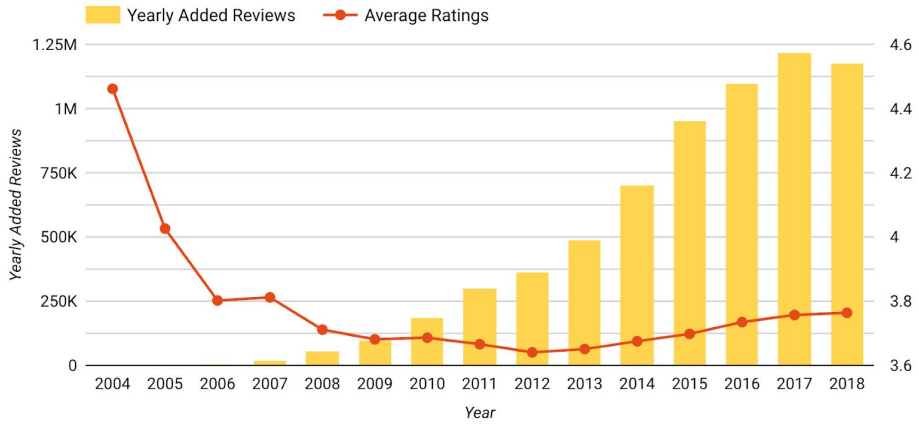
- Reviews
  Here reviews are displayed at a monthly interval. 12 most useful reviews, one for each month in the latest year are selected to give a detailed description of how the business reputation changes along the year.

- Users
  For analysis towards users, I mainly studied on these topics:
  - How active are the users?
  - What does the most ‘experienced’ user look like?
  - Is there any suspicious user?

  A quick visualization of user’s activity and rating distribution:

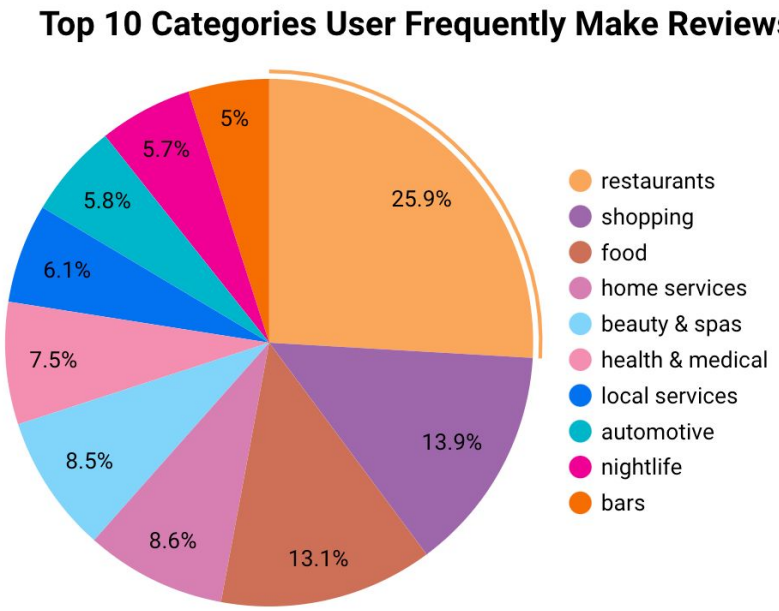


What within expectation are: nearly half of the users have made under 100 reviews(me myself is a lazy person, I just read but not commented), and more than half stars are greater than 3, with very few under 2(customers are kind). What out of expectation is that there are >5% users made over 1000 reviews!



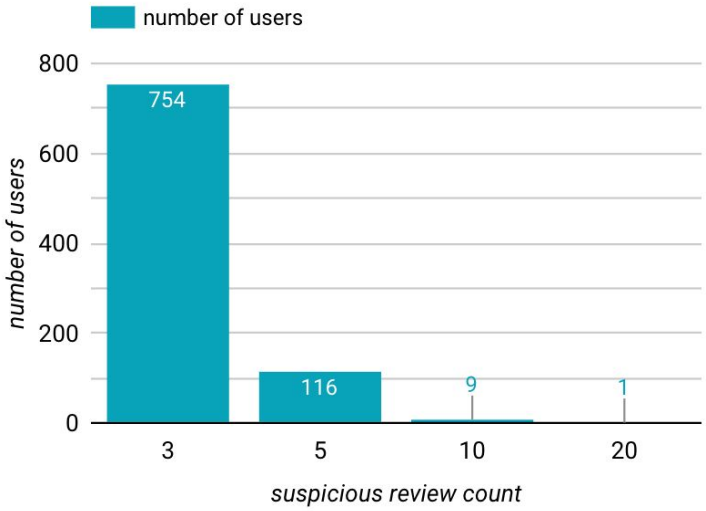
Let’s revisit this graph and take a look at the yearly average stars all users made.
We can see as the number of reviews increase, the average star tends to stabilize around 3.7, which is consistent with what we found in the pie charts.

Move on, for those users who made so many reviews, I was really curious about what kind of business they are interested in, so among all users, I extracted the top 20 users with the most reviews, and counted all the business categories that they've reviewed. Here are the 10 most frequent categories:



No surprise, 'restaurants' and 'food' together take ~40%(although still lower than I expected). Food recommendation is what users care more. But with such a good user base, I think if Yelp can guide users to give reviews on other important business, like health&medical, it can have profound social impact.

Just like what I did in the business analysis section, suspicious users were identified as users that make multiple reviews with the same rating on one business within one year. With different thresholds I got:



A low threshold like 3 or 5 doesn't necessarily mean the user is suspicious, maybe he is a huge fan of this restaurant, but if one user makes more than 20 reviews on one business, then it's very likely that he is a bot or he is bribed.
Meanwhile let's take a look at the most active users:

Average Fans	Average Reviews	Average Useful
1.47	22.29	40.53

Review #	Fan #	Useful Per Review	User Count ▼
1000	100	5	384
2000	200	6	66
3000	300	8	6
5000	500	10	1

The above figure first shows the average fans/reviews/useful per review each user owns. Then it shows the number of 'popular users' with different thresholds.

With these figures, I further looked into the impact of these active users. Take the 384 active users who have more than 1000 reviews, 5 useful per review and 100 fans, and their reviews cover 2817 businesses out of 192609 businesses in total. There are 179 out of the 2817 businesses that have more than 1000 reviews, meanwhile there are 348 businesses that have no less reviews in total. These users are merely 0.02% of all users, but they've reviewed more than half of the popular businesses. From here we can see the huge impact popular users can bring for a business.

Challenges

- What is really worth analyzing?
  I felt the hardest part for a data analysis project, rather than how to realize, is what to analyze. When I first got this dataset I had too many ideas, but when I reviewed each idea, I found those were not really 'analysis', but just visualization, or utilize those data to implement a feature. An analysis should be a question at first, and then process the data to jump to an answer or a conclusion. Then conclusion should be meaningful, e.g. can give useful suggestions on what the company can do to grow. Even now I am not sure if what I did belongs to data analysis.

- Huge cost on queries with parameters.
  At first I felt lucky that the dataset was clean and didn't require much pre-processing, but later when I started analyzing, I found that some queries needed to join multiple tables, and the cost was expensive.
  For example, I had a query that extracted the three most useful reviews for one business in a specific year, and during one execution, business and reviews tables were joined, and reviews had to be filtered and partitioned based on years. The cost of one execution is ~4GB.
  What's worse was that the queried business name was set as a customized user input in Data Studio, which means caching is also not a good option, since every time the business name can change. I've also partitioned the reviews table by year, but it didn't help. The most time-consuming part was the join operation, so the only optimization I could come up with is to create a temporary table that connects business and reviews.

- Complicated transformation between raw data and data to display on report
  Even if BigQuery has provided a very-well organized data payload, many subtle changes are still required for better visualization. This may not be a big technical obstacle, but it WAS time-consuming.
  Some fields are stored as repeated or record types in Bigquery, but these types are hard for Data Studio to display, so I had to figure out a workable payload structure. For example I had to concatenate an array of strings to a single string and split it later, or I have to reformat the geographical fields in BigQuery so that Data Studio geo tools can correctly parse these values. I also had to decide where to conduct some calculations, e.g. it's easy to get a sum of number of reviews for each year in BigQuery, but to get a running sum it's more convenient to do it on Data Studio.

- Reorganization of data
  As the analysis went on, more ideas jumped into my mind, some of which required further processing of the original tables to realize a good performance (for example for the query in the first bullet point, it will be helpful to have an auxiliary table, but I haven't done that yet).

Takeaways

- Yelp
  Yelp's definitely a success, yet it still has much potential to grow. Some suggestions on their business plans:

1) Broaden business scopes to other fields, like shopping, beauty, health and medical, etc.
2) Keep increasing user interactions.

- Business

1) Yelp's a good platform to increase visibility to the public.
2) Value suggestions and complaints given by hot reviews, because people care.
3) Even though the possibility is very low, pay attention to competitors' deliberate bad reviews.

- Users

1) Yelp's users are generally 'kind', with an average rating approaching 4.
2) Yelp's users are 'real' users.
3) Popular(or active) users have a huge impact on businesses' popularity.

Future Work

- If more data: more reasonable geographical distribution of businesses. All the analysis above can be more precise.
- Recommendations: based on the analysis, it will be interesting to generate recommendations for users based on their review history.
- Natural Language Processing(too hard for me): based on the review/tip text, summarize a few keywords to describe the business, or generate suggestions for businesses to improve itself.

Why BigQuery

- GCP integration
  Supported by Google Cloud Platform, BigQuery can easily integrate with other services like GCS and Google Data Studio. I don't have to worry about configuration settings between different apps.
- Good scalability, high efficiency and low cost
  Processing gigabyte-sized files consumes much resources on a local laptop. From data uploading, to data storage, to processing queries, to scale up and monitoring, Google takes care of everything. Plus, a few GB is like nothing for Google(thus nearly no cost). Why not?
  A few good practices helped me save even more quota. For example, retrieve only the columns I want, or partition the table based on time. And the query time is just within a few seconds.
- Rich Tools & Simple CLI
  BigQuery has many useful features to improve performance. Uploading/creating tables can be done within one command via Google cloud interactive CLI. It is also convenient to create tables in BigQuery's UI, and it can automatically detect the schemas.
  For processing queries, it's easy to query between different tables. It also allows caching costly query's result to save you quota. I can also quickly create a temporary table/google sheet to store the results, or just export the data directly to Data Studio, all through one mouse click.
  One thing I find interesting is working with BigQuery's GIS data. Users can create a geo field in BigQuery with different data formats, e.g. lat/lon pairs. After formalizing data to the geo type, BQ then provides a rich set of functions to play with it. For example, I converted the lat/lon fields in my business table to geo field and joined another urban_area table in google public dataset to allocate businesses in different cities, based on which I carried out a series of analysis on geographics.
- Concise UI
  BigQuery's UI provides a big query window with auto-formatting. It also helps detect syntax error and estimate potential cost. Query results are displayed elegantly, with a few options to further decide the display format. Users can also save queries that are called frequently as well.

Why Data Studio

- Rich data sources
  Data Studio accepts almost all kinds of Google data sources: BigQuery, Google Analytics, Google sheets, etc. Also, it allows large amounts of data sources in one report(it starts pricing when the number exceeds the threshold, though).
- Quick to start
  Data Studio is very user-friendly. I just watched a tutorial video before starting. Google also provides tons of docs and tutorial videos.
- Fancy Visualization & Flexible Data Handling
  Data Studio provides all kinds of graphs you can imagine. One thing that Data Studio impresses me is the ability to support custom queries. For example in my report, if I want to see data for different businesses, I can merely type in the business name I want, and it will modify the sql query to fetch the matched results and display. Data Studio also provides some features to calculate new fields based on the imported data sources.

Sample Queries

Get 3 most popular categories in each city:

```
WITH
  urban AS (
  SELECT
    name,
    urban_area_geom AS geo
  FROM
    `bigquery-public-data.geo_us_boundaries.urban_areas`
  LIMIT
    1000),
  business_geo AS (
  SELECT
    name,
    categories,
    business_id,
    ST_GeogPoint(longitude,
      latitude) AS geo
  FROM
    `yelp-food-hunter.yelp_data_set.business`)
SELECT
  *
FROM (
  SELECT
    *,
    ROW_NUMBER() OVER (PARTITION BY name ORDER BY count DESC) AS rank
  FROM (
    SELECT
      name,
      LOWER(TRIM(category)) AS category,
      COUNT(business_id) AS count
    FROM (
      SELECT
        urban.name,
        business_geo.categories,
        business_geo.business_id
      FROM
        urban
      JOIN
        business_geo
      ON
        ST_COVEREDBY(business_geo.geo,
          urban.geo)),
      UNNEST(categories) AS category
    GROUP BY
      1,
      2))
WHERE
  rank <= 3;
```

- Get suspicious users for one business

```
WITH
  a AS (
  WITH
    tmp AS (
    SELECT
      user_id,
      business_id,
      COUNT(business_id) AS count,
      stddev(stars) AS star_dev,
      EXTRACT(YEAR
      FROM
        date) AS year
    FROM
      `yelp-food-hunter.yelp_data_set.reviews`
    GROUP BY
      1,
      2,
      year)
  SELECT
    COUNT(user_id) AS suspicious_user_count,
    business_id
  FROM
    tmp
  WHERE
    count > 3
    AND star_dev = 0
    AND business_id = "iCQpiavjjPzJ5_3gPD5Ebg"
  GROUP BY
    year,
    business_id
  ORDER BY
    1 DESC),
  b AS (
  SELECT
    business_id,
    COUNT(review_id) AS review_count,
    COUNT(DISTINCT user_id) AS user_count,
    ROUND(SUM(useful)/COUNT(review_id),1) AS useful_per_review,
    ROUND(stddev(stars), 1) AS star_dev,
    (MAX(EXTRACT(year
        FROM
          date))-MIN(EXTRACT(year
        FROM
          date))) AS on_yelp
  FROM
    `yelp-food-hunter.yelp_data_set.reviews`
  WHERE
    business_id = "iCQpiavjjPzJ5_3gPD5Ebg"
  GROUP BY
    1)
SELECT
  b.*,
  ifnull(a.suspicious_user_count,
    0) AS suspicious_user_count
FROM
  b
LEFT JOIN
  a
ON
  b.business_id = a.business_id
```

- Get reviews with the most useful tags in the latest 12 months

```
SELECT
  text,
  useful,
  stars,
  month,
  year
FROM (
  SELECT
    *,
    ROW_NUMBER() OVER (PARTITION BY month ORDER BY year DESC, useful DESC) AS rank
  FROM (
    SELECT
      text,
      useful,
      stars,
      EXTRACT(month
      FROM
        date) AS month,
      EXTRACT(year
      FROM
        date) AS year
    FROM
      `yelp-food-hunter.yelp_data_set.reviews`
    WHERE
      business_id = @id
    ORDER BY
      4,
      useful DESC ) )
WHERE
  rank = 1
ORDER BY
  year DESC,
  month DESC
```