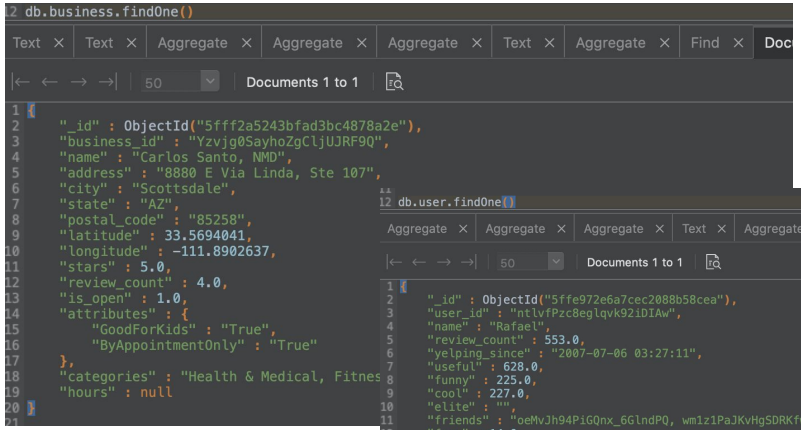
Exploratory Data Analysis of Yelp Data

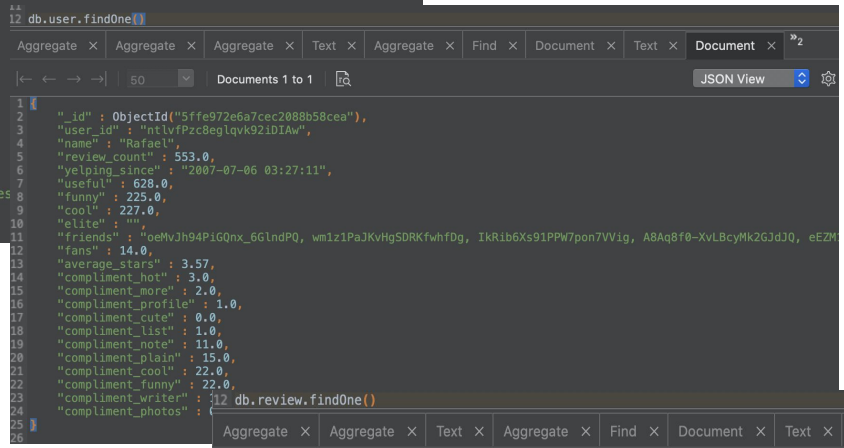
Sophie Wang, Elyse Cheung-Sutton, Aneri Dand, Vaishnavi Kashyap, Qianyun(April) Li

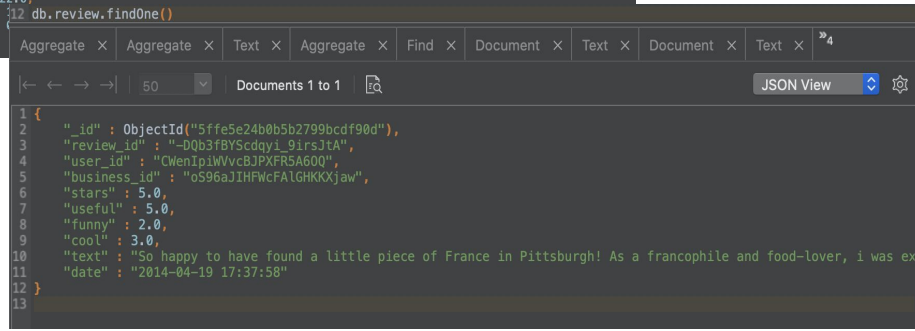
Background

5 collections:

1. Business
2. Users
3. Check-ins
4. Reviews
5. Tips





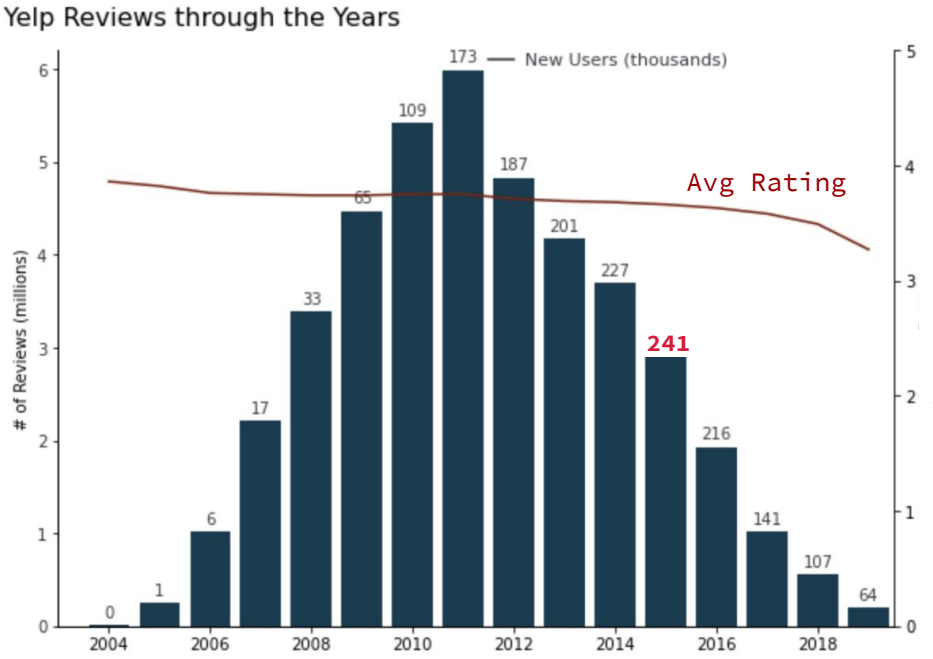


Initial EDA:

1. Only 37 states including Canadian states
2. 2011 had the most reviews

How many Reviews and new Users between 2004-2019?

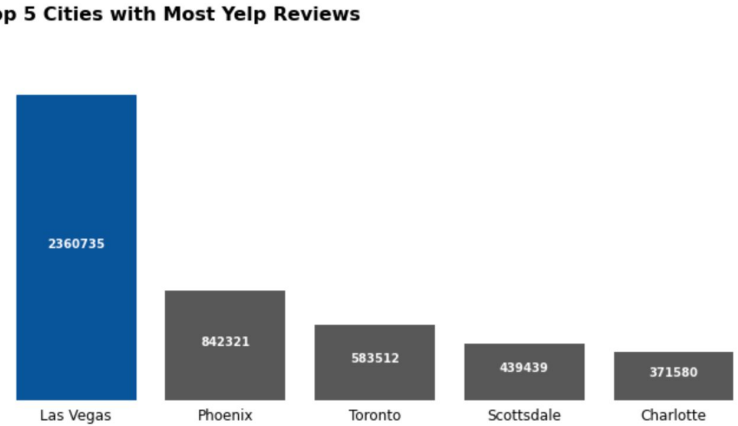
- 2015: most new Users
- 2011: most reviews
- Average review rating constant
- Expect 2019 to have most reviews and new users given that Yelp is always growing



Top 5 Cities w/ Most Reviews & Categories Most Reviewed in Top City

- Las Vegas ranked #1 with 2,360,735 reviews - not surprising

- Different versions of “Mexican Restaurant” and “Nail Salon, Beauty” made up the Top 5 Categories reviewed in Las Vegas - a bit surprising, would have thought “Bars/Clubs” would be in the top 5 reviewed

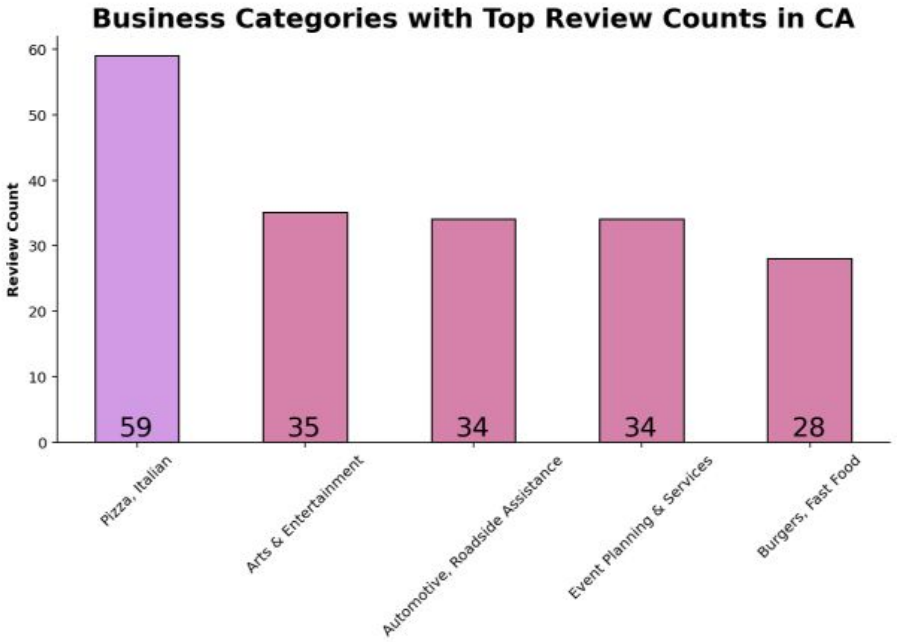



```
{
    "category" : "Mexican, Restaurants",
    "num_reviews" : 143.0
}
{
    "category" : "Restaurants, Mexican",
    "num_reviews" : 133.0
}
{
    "category" : "Food, Coffee & Tea",
    "num_reviews" : 114.0
}
{
    "category" : "Nail Salons, Beauty & Spas",
    "num_reviews" : 108.0
}
{
    "category" : "Beauty & Spas, Nail Salons",
    "num_reviews" : 107.0
```

Which business categories in California have the most review counts?

- The business categories with top review counts are Italian food, entertainment, automotive, event Planning and fast food in California (based on this dataset).

- Caveat: This dataset does not represent the comprehensive yelp data within US. (Number of business in CA is a lot smaller than other states). So it’s hard to say in reality which category in CA has the most reviews. Also high review count does not equal high popularity.



What do people complain most about restaurants?

Top 5 Common Nouns in Negative Reviews

Food	Service	Place	Time	Order

- For restaurants, no wonder terrible food is the most frequent reason for negative reviews.

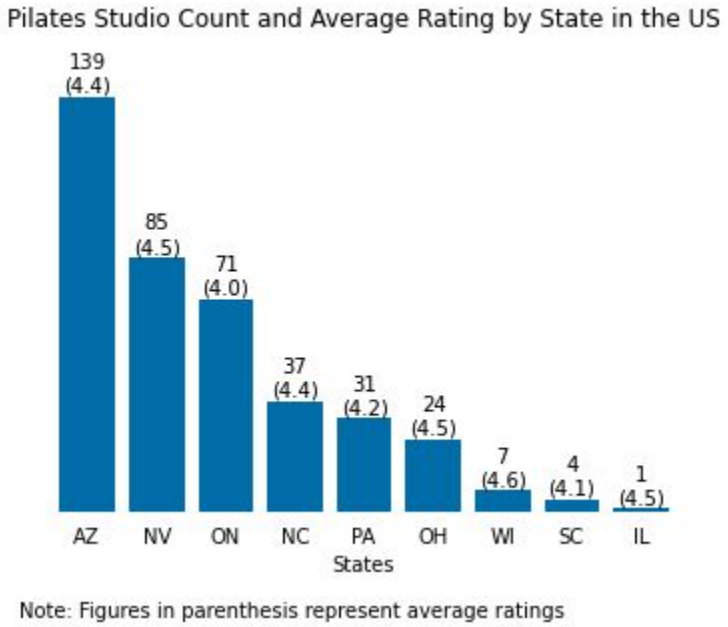
- Besides, restaurants should also pay attention to service, place, waiting time and order service in case of negative reviews.

Which states have the highest number of pilates studios?

- Based on EDA, it appears that Arizona has the highest number of Pilates studios.

- However, the veracity of the trend shown in the adjacent graph comes into question as the Yelp dataset contains information only about 9 US states.

- California and Texas (US states with the highest number of gyms*) are not included in the data.



*Source: https://www.exercise.com/learn/ten-fittest-states-in-the-us/ https://www.ibisworld.com/industry-statistics/number-of-businesses/gym-health-fitness-clubs-united-states/

Data Pipeline Efficiency

- From pymongo (not distributed computing setting) to spark, we achieve 95% runtime reduction.

- In Spark, using techniques to boost efficiency we were able to run and execute code in ~40% less time.

Key Takeaways

- Running multiple jupyter notebooks on the same cluster was very slow owing to memory constraints

- Having the flexibility to control the number of partitions helps improve query efficiency

- Large data does not always translate to better / accurate insights