# Yelp Data Analysis

Group 5: Yi-Yun Liao, Hongfei Chen, Hsin Yeh, Xiaotian He

## 1. Abstract

In this project, we analyzed the restaurant rating and review datasets on Yelp to provide insights into the success patterns of popular restaurants in the US. We utilized Hadoop for data cleaning and Hive for analyzing Yelp reviews of successful US restaurants. By integrating this with US demographic data, we can gain a comprehensive understanding of the factors that contribute to a restaurant's success.

## 2. Introduction

In the analysis, we used only the data from users with an average star rating higher than 3 as 'Good Users', as their reviews may be more reliable and informative, and businesses with 50 or more reviews, as this can help ensure a sufficient sample size for analysis and reduce the impact of outliers or biased reviews.

In the highly competitive restaurant industry, understanding the factors that contribute to a successful restaurant can provide a competitive advantage. By analyzing Yelp data, this project can provide insights into user behavior, preferences, and opinions, which can help restaurant owners and managers make data-driven decisions to improve their business. This research can be beneficial for a wide range of users, including restaurant owners and managers, investors, researchers, and analysts who are interested in understanding the factors that influence restaurant success. Restaurant owners and managers can benefit from this analysis by gaining insights into the factors that contribute to successful restaurants, such as user ratings and reviews, and can make informed decisions to improve their business. Investors can use this analysis to identify potential profitable restaurant investments. Researchers and analysts can use this data to study trends and patterns in the restaurant industry.

## 3. Pre-processing

### 3.1. Data Source

Our original datasets contain three business, review and user datasets from Yelp and one demographic information dataset from NYC OpenData.

**a. Dataset: Yelp_academic_dataset_business**

- Description: Businesses listed on Yelp
- Size: 118.86 MB

**b. Dataset: Yelp_academic_dataset_user**

- Description: User information from Yelp
- Size: 3.36 GB

**c. Dataset: Yelp_academic_dataset_review**

- Description: Millions of reviews from Yelp
- Size: 5.34 GB

**d. Dataset: Demographic Statistics By Zip Code**

- Description: Demographic information from NYC OpenData
- Size: 29KB

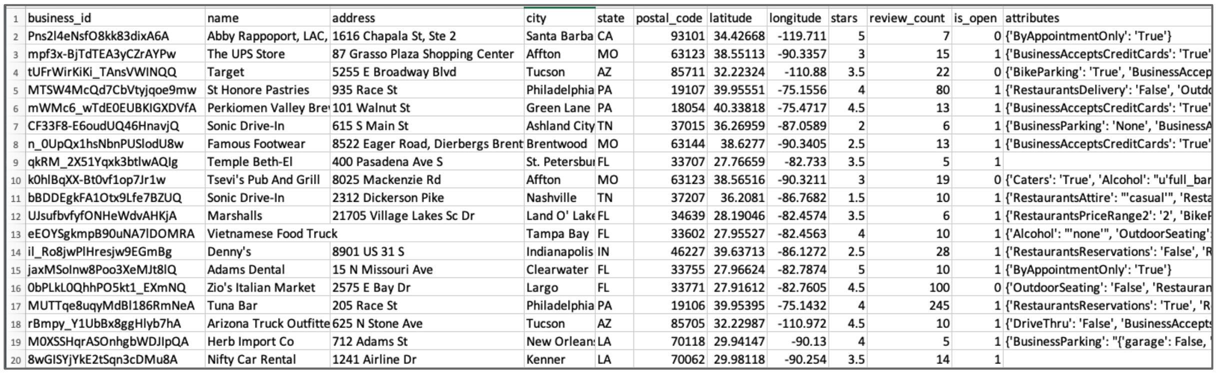
## 3.2. Data Cleaning

The Yelp dataset contains a large number of attributes, but not all of them may be relevant for the analysis. Therefore, it is essential to identify the relevant features and ensure that they are included in the analysis. The demographic dataset contains gender, race, nationality and public assistance information with both actual numbers and percentages, given that the total people participated in the survey are different in different jurisdictions, we decided to use only the percentages data to ensure the comparability between different regions.

In this analysis, we use Mapreduce to clean the data and select relevant attributes on NYU Dataproc Platform.

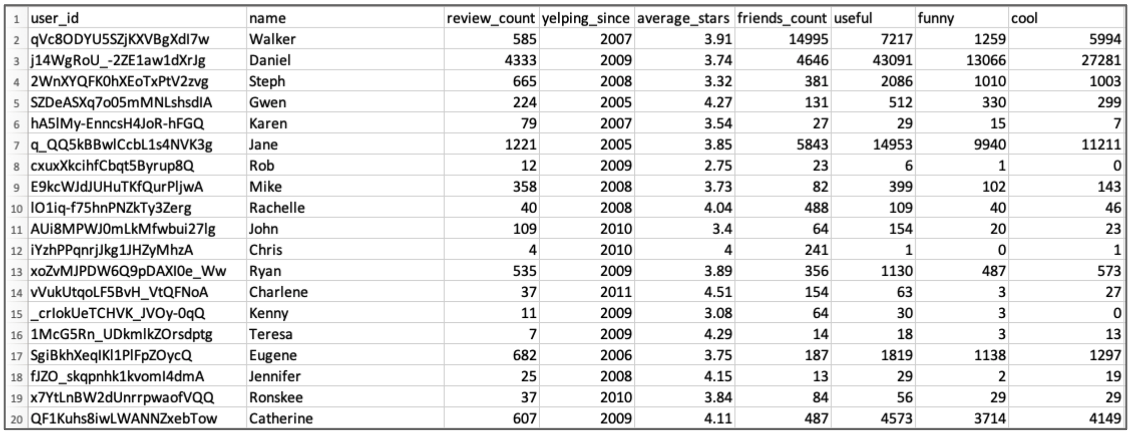
## 3.3. Data Sample

**a. Dataset: Yelp_academic_dataset_business**



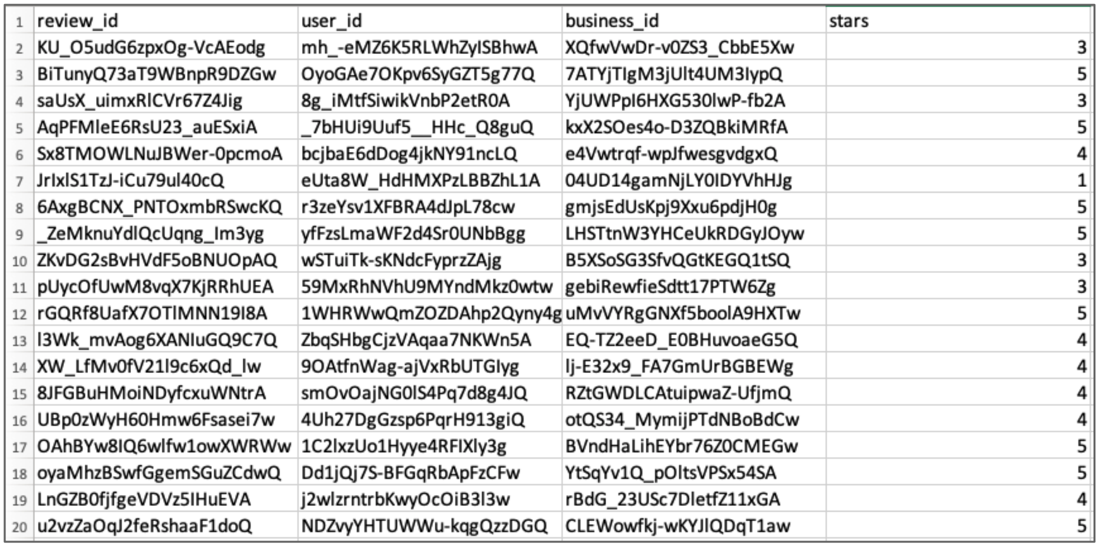
| 1 | business_id | name | address | city | state | postal_code | latitude | longitude | stars | review_count | is_open | attributes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | Pns2l4eNsfO8kk83dixA6A | Abby Rappoport, LAC, | 1616 Chapala St, Ste 2 | Santa Barba | CA | 93101 | 34.42668 | -119.711 | 5 | 7 | 0 | {'ByAppointmentOnly': 'True'} |
| 3 | mpf3x-BjTdTEA3yCZrAYPw | The UPS Store | 87 Grasso Plaza Shopping Center | Affton | MO | 63123 | 38.55113 | -90.3357 | 3 | 15 | 1 | {'BusinessAcceptsCreditCards': 'True' |
| 4 | tUFrWirKiKi_TAnsVWINQQ | Target | 5255 E Broadway Blvd | Tucson | AZ | 85711 | 32.22324 | -110.88 | 3.5 | 22 | 0 | {'BikeParking': 'True', 'BusinessAccep |
| 5 | MTSW4McQd7CbVtyjqoe9mw | St Honore Pastries | 935 Race St | Philadelphia | PA | 19107 | 39.95551 | -75.1556 | 4 | 80 | 1 | {'RestaurantsDelivery': 'False', 'Outdo |
| 6 | mWMc6_wTdE0EUBKIGXDVfA | Perkiomen Valley Brev | 101 Walnut St | Green Lane | PA | 18054 | 40.33818 | -75.4717 | 4.5 | 13 | 1 | {'BusinessAcceptsCreditCards': 'True' |
| 7 | CF33F8-E6oudUQ46HnavjQ | Sonic Drive-In | 615 S Main St | Ashland City | TN | 37015 | 36.26959 | -87.0589 | 2 | 6 | 1 | {'BusinessParking': 'None', 'BusinessA |
| 8 | n_0UpQx1hsNbnPUSlodU8w | Famous Footwear | 8522 Eager Road, Dierbergs Brent | Brentwood | MO | 63144 | 38.6277 | -90.3405 | 2.5 | 13 | 1 | {'BusinessAcceptsCreditCards': 'True' |
| 9 | qkRM_2X51Yqxk3btlwAQlg | Temple Beth-El | 400 Pasadena Ave S | St. Petersbur | FL | 33707 | 27.76659 | -82.733 | 3.5 | 5 | 1 | |
| 10 | k0hlBqXX-Bt0vf1op7Jr1w | Tsevi's Pub And Grill | 8025 Mackenzie Rd | Affton | MO | 63123 | 38.56516 | -90.3211 | 3 | 19 | 0 | {'Caters': 'True', 'Alcohol': "u'full_bar |
| 11 | bBDDEgkFA1Otx9Lfe7BZUQ | Sonic Drive-In | 2312 Dickerson Pike | Nashville | TN | 37207 | 36.2081 | -86.7682 | 1.5 | 10 | 1 | {'RestaurantsAttire': "'casual'", 'Resta |
| 12 | UJsufbvfyfONHeWdvAHKjA | Marshalls | 21705 Village Lakes Sc Dr | Land O' Lake | FL | 34639 | 28.19046 | -82.4574 | 3.5 | 6 | 1 | {'RestaurantsPriceRange2': '2', 'BikeP |
| 13 | eEOYSgkmpB90uNA7lDOMRA | Vietnamese Food Truck | | Tampa Bay | FL | 33602 | 27.95527 | -82.4563 | 4 | 10 | 1 | {'Alcohol': "'none'", 'OutdoorSeating': |
| 14 | il_Ro8jwPlHresjw9EGmBg | Denny's | 8901 US 31 S | Indianapolis | IN | 46227 | 39.63713 | -86.1272 | 2.5 | 28 | 1 | {'RestaurantsReservations': 'False', 'R |
| 15 | jaxMSoInw8Poo3XeMJt8lQ | Adams Dental | 15 N Missouri Ave | Clearwater | FL | 33755 | 27.96624 | -82.7874 | 5 | 10 | 1 | {'ByAppointmentOnly': 'True'} |
| 16 | 0bPLkL0QhhPO5kt1_EXmNQ | Zio's Italian Market | 2575 E Bay Dr | Largo | FL | 33771 | 27.91612 | -82.7605 | 4.5 | 100 | 0 | {'OutdoorSeating': 'False', 'Restauran |
| 17 | MUTTqe8uqyMdBl186RmNeA | Tuna Bar | 205 Race St | Philadelphia | PA | 19106 | 39.95395 | -75.1432 | 4 | 245 | 1 | {'RestaurantsReservations': 'True', 'R |
| 18 | rBmpy_Y1UbBx8ggHlyb7hA | Arizona Truck Outfitte | 625 N Stone Ave | Tucson | AZ | 85705 | 32.22987 | -110.972 | 4.5 | 10 | 1 | {'DriveThru': 'False', 'BusinessAccepts |
| 19 | M0XSSHqrASOnhgbWDJIpQA | Herb Import Co | 712 Adams St | New Orleans | LA | 70118 | 29.94147 | -90.13 | 4 | 5 | 1 | {'BusinessParking': "{'garage': False, ' |
| 20 | 8wGISYjYkE2tSqn3cDMu8A | Nifty Car Rental | 1241 Airline Dr | Kenner | LA | 70062 | 29.98118 | -90.254 | 3.5 | 14 | 1 | |

## b. Dataset: Yelp_academic_dataset_user



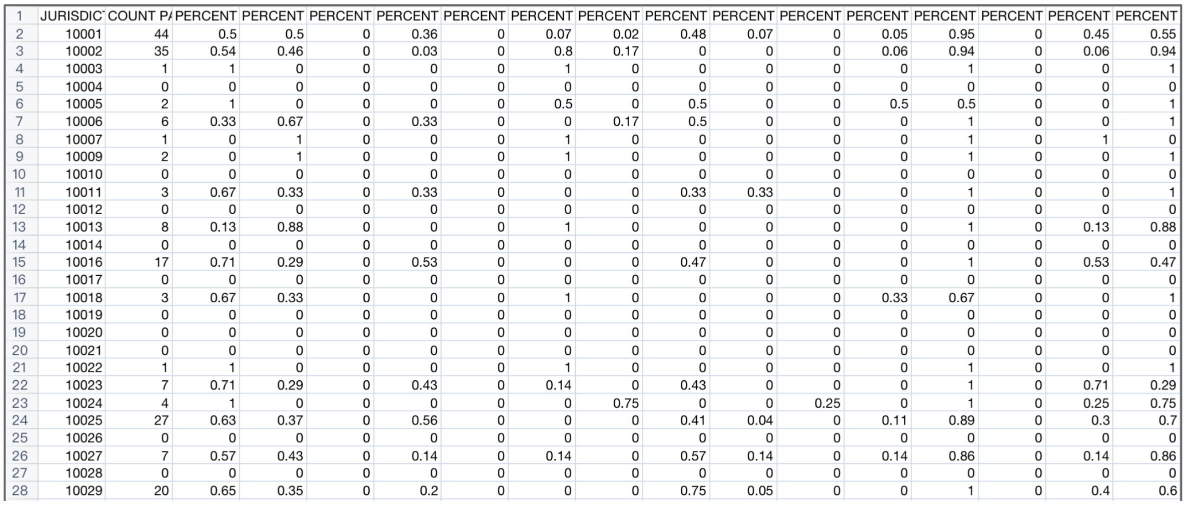
| user_id | name | review_count | yelping_since | average_stars | friends_count | useful | funny | cool |
|---|---|---|---|---|---|---|---|---|
| qVc8ODYU5SZjKXVBgXdI7w | Walker | 585 | 2007 | 3.91 | 14995 | 7217 | 1259 | 5994 |
| j14WgRoU_-2ZE1aw1dXrJg | Daniel | 4333 | 2009 | 3.74 | 4646 | 43091 | 13066 | 27281 |
| 2WnXYQFK0hXEoTxPtV2zvg | Steph | 665 | 2008 | 3.32 | 381 | 2086 | 1010 | 1003 |
| SZDeASXq7o05mMNLshsdIA | Gwen | 224 | 2005 | 4.27 | 131 | 512 | 330 | 299 |
| hA5lMy-EnncsH4JoR-hFGQ | Karen | 79 | 2007 | 3.54 | 27 | 29 | 15 | 7 |
| q_QQ5kBBwlCcbL1s4NVK3g | Jane | 1221 | 2005 | 3.85 | 5843 | 14953 | 9940 | 11211 |
| cxuxXkcihfCbqt5Byrup8Q | Rob | 12 | 2009 | 2.75 | 23 | 6 | 1 | 0 |
| E9kcWJdJUHuTKfQurPljwA | Mike | 358 | 2008 | 3.73 | 82 | 399 | 102 | 143 |
| lO1iq-f75hnPNZkTy3Zerg | Rachelle | 40 | 2008 | 4.04 | 488 | 109 | 40 | 46 |
| AUi8MPWJ0mLkMfwbui27lg | John | 109 | 2010 | 3.4 | 64 | 154 | 20 | 23 |
| iYzhPPqnrjJkg1JHZyMhzA | Chris | 4 | 2010 | 4 | 241 | 1 | 0 | 1 |
| xoZvMJPDW6Q9pDAXI0e_Ww | Ryan | 535 | 2009 | 3.89 | 356 | 1130 | 487 | 573 |
| vVukUtqoLF5BvH_VtQFNoA | Charlene | 37 | 2011 | 4.51 | 154 | 63 | 3 | 27 |
| _crIokUeTCHVK_JVOy-0qQ | Kenny | 11 | 2009 | 3.08 | 64 | 30 | 3 | 0 |
| 1McG5Rn_UDkmlkZOrsdptg | Teresa | 7 | 2009 | 4.29 | 14 | 18 | 3 | 13 |
| SgiBkhXeqIKl1PlFpZOycQ | Eugene | 682 | 2006 | 3.75 | 187 | 1819 | 1138 | 1297 |
| fJZO_skqpnhk1kvomI4dmA | Jennifer | 25 | 2008 | 4.15 | 13 | 29 | 2 | 19 |
| x7YtLnBW2dUnrrpwaofVQQ | Ronskee | 37 | 2010 | 3.84 | 84 | 56 | 29 | 29 |
| QF1Kuhs8iwLWANNZxebTow | Catherine | 607 | 2009 | 4.11 | 487 | 4573 | 3714 | 4149 |

## c. Dataset: Yelp_academic_dataset_review



| review_id | user_id | business_id | stars |
|---|---|---|---|
| KU_O5udG6zpxOg-VcAEodg | mh_-eMZ6K5RLWhZyISBhwA | XQfwVwDr-v0ZS3_CbbE5Xw | 3 |
| BiTunyQ73aT9WBnpR9DZGw | OyoGAe7OKpv6SyGZT5g77Q | 7ATYjTIgM3jUlt4UM3IypQ | 5 |
| saUsX_uimxRlCVr67Z4Jig | 8g_iMtfSiwikVnbP2etR0A | YjUWPpI6HXG530lwP-fb2A | 3 |
| AqPFMleE6RsU23_auESxiA | _7bHUi9Uuf5__HHc_Q8guQ | kxX2SOes4o-D3ZQBkiMRfA | 5 |
| Sx8TMOWLNuJBWer-0pcmoA | bcjbaE6dDog4jkNY91ncLQ | e4Vwtrqf-wpJfwesgvdgxQ | 4 |
| JrIxlS1TzJ-iCu79ul40cQ | eUta8W_HdHMXPzLBBZhL1A | 04UD14gamNjLY0IDYVhHJg | 1 |
| 6AxgBCNX_PNTOxmbRSwcKQ | r3zeYsv1XFBRA4dJpL78cw | gmjsEdUsKpj9Xxu6pdjH0g | 5 |
| _ZeMknuYdlQcUqng_Im3yg | yfFzsLmaWF2d4Sr0UNbBgg | LHSTtnW3YHCeUkRDGyJOyw | 5 |
| ZKvDG2sBvHVdF5oBNUOpAQ | wSTuiTk-sKNdcFyprzZAjg | B5XSoSG3SfvQGtKEGQ1tSQ | 3 |
| pUycOfUwM8vqX7KjRRhUEA | 59MxRhNVhU9MYndMkz0wtw | gebiRewfieSdtt17PTW6Zg | 3 |
| rGQRf8UafX7OTlMNN19I8A | 1WHRWwQmZOZDAhp2Qyny4g | uMvVYRgGNXf5boolA9HXTw | 5 |
| l3Wk_mvAog6XANIuGQ9C7Q | ZbqSHbgCjzVAqaa7NKWn5A | EQ-TZ2eeD_E0BHuvoaeG5Q | 4 |
| XW_LfMv0fV21l9c6xQd_lw | 9OAtfnWag-ajVxRbUTGIyg | lj-E32x9_FA7GmUrBGBEWg | 4 |
| 8JFGBuHMoiNDyfcxuWNtrA | smOvOajNG0lS4Pq7d8g4JQ | RZtGWDLCAtuipwaZ-UfjmQ | 4 |
| UBp0zWyH60Hmw6Fsasei7w | 4Uh27DgGzsp6PqrH913giQ | otQS34_MymijPTdNBoBdCw | 4 |
| OAhBYw8IQ6wlfw1owXWRWw | 1C2lxzUo1Hyye4RFIXly3g | BVndHaLihEYbr76Z0CMEGw | 5 |
| oyaMhzBSwfGgemSGuZCdwQ | Dd1jQj7S-BFGqRbApFzCFw | YtSqYv1Q_pOltsVPSx54SA | 5 |
| LnGZB0fjfgeVDVz5IHuEVA | j2wlzrntrbKwyOcOiB3l3w | rBdG_23USc7DletfZ11xGA | 4 |
| u2vzZaOqJ2feRshaaF1doQ | NDZvyYHTUWWu-kqgQzzDGQ | CLEWowfkj-wKYJlQDqT1aw | 5 |

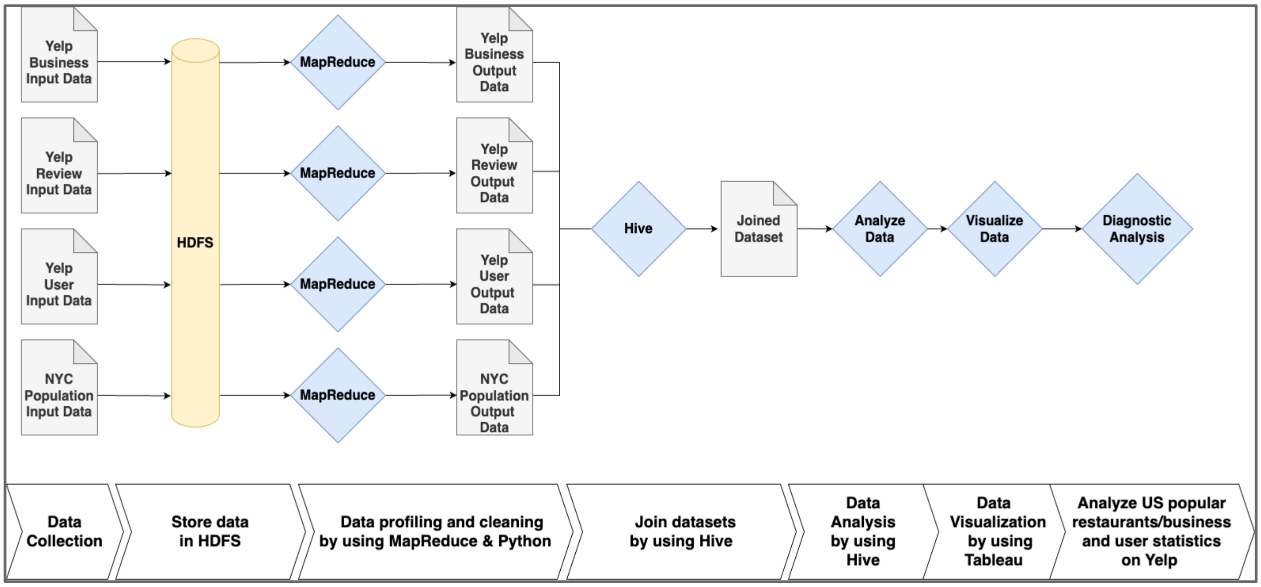
### d. Dataset: Demographic Statistics By Zip Code



| 1 | JURISDICT | COUNT PA | PERCENT | PERCENT | PERCENT | PERCENT | PERCENT | PERCENT | PERCENT | PERCENT | PERCENT | PERCENT | PERCENT | PERCENT | PERCENT | PERCENT | PERCENT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 10001 | 44 | 0.5 | 0.5 | 0 | 0.36 | 0 | 0.07 | 0.02 | 0.48 | 0.07 | 0 | 0.05 | 0.95 | 0 | 0.45 | 0.55 |
| 3 | 10002 | 35 | 0.54 | 0.46 | 0 | 0.03 | 0 | 0.8 | 0.17 | 0 | 0 | 0 | 0.06 | 0.94 | 0 | 0.06 | 0.94 |
| 4 | 10003 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 5 | 10004 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 10005 | 2 | 1 | 0 | 0 | 0 | 0 | 0.5 | 0 | 0.5 | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 1 |
| 7 | 10006 | 6 | 0.33 | 0.67 | 0 | 0.33 | 0 | 0 | 0.17 | 0.5 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 8 | 10007 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 9 | 10009 | 2 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 10 | 10010 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 10011 | 3 | 0.67 | 0.33 | 0 | 0.33 | 0 | 0 | 0 | 0.33 | 0.33 | 0 | 0 | 1 | 0 | 0 | 1 |
| 12 | 10012 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 10013 | 8 | 0.13 | 0.88 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0.13 | 0.88 |
| 14 | 10014 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 10016 | 17 | 0.71 | 0.29 | 0 | 0.53 | 0 | 0 | 0 | 0.47 | 0 | 0 | 0 | 1 | 0 | 0.53 | 0.47 |
| 16 | 10017 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 10018 | 3 | 0.67 | 0.33 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0.33 | 0.67 | 0 | 0 | 1 |
| 18 | 10019 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 10020 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | 10021 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | 10022 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 22 | 10023 | 7 | 0.71 | 0.29 | 0 | 0.43 | 0 | 0.14 | 0 | 0.43 | 0 | 0 | 0 | 1 | 0 | 0.71 | 0.29 |
| 23 | 10024 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0.75 | 0 | 0 | 0.25 | 0 | 1 | 0 | 0.25 | 0.75 |
| 24 | 10025 | 27 | 0.63 | 0.37 | 0 | 0.56 | 0 | 0 | 0 | 0.41 | 0.04 | 0 | 0.11 | 0.89 | 0 | 0.3 | 0.7 |
| 25 | 10026 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 26 | 10027 | 7 | 0.57 | 0.43 | 0 | 0.14 | 0 | 0.14 | 0 | 0.57 | 0.14 | 0 | 0.14 | 0.86 | 0 | 0.14 | 0.86 |
| 27 | 10028 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | 10029 | 20 | 0.65 | 0.35 | 0 | 0.2 | 0 | 0 | 0 | 0.75 | 0.05 | 0 | 0 | 1 | 0 | 0.4 | 0.6 |

# 4. Data Analysis

## 4.1. Analysis Diagram

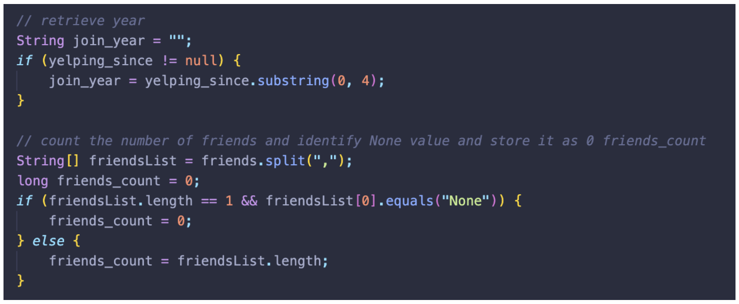

Yelp Business Input Data
Yelp Review Input Data
Yelp User Input Data
NYC Population Input Data
HDFS
MapReduce
MapReduce
MapReduce
MapReduce
Yelp Business Output Data
Yelp Review Output Data
Yelp User Output Data
NYC Population Output Data
Hive
Joined Dataset
Analyze Data
Visualize Data
Diagnostic Analysis
Data Collection
Store data in HDFS
Data profiling and cleaning by using MapReduce & Python
Join datasets by using Hive
Data Analysis by using Hive
Data Visualization by using Tableau
Analyze US popular restaurants/business and user statistics on Yelp

## 4.2. Code Challenge

### a. Extract/Calculate/Store useful values in Yelp User.json

The yelp_since in the user dataset contains complete timestamp yyyy-mm-dd hh:mm:ss, which has redundant information we don't need for further data analysis in Hive or Tableau. Besides, the friends in the user dataset contains user ids for all friends which are useful for data analysis or data visualization.

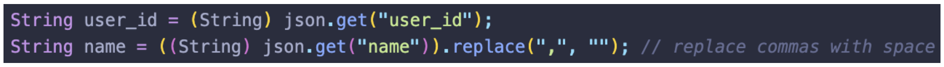
Our solution is to substring yelp_since and only extract year value. Create a new field friends_count to store the number of friends for each user instead of the user ids list for their friends.



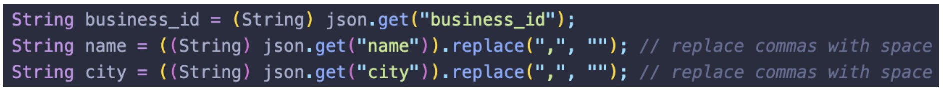
```
// retrieve year
String join_year = "";
if (yelping_since != null) {
    join_year = yelping_since.substring(0, 4);
}

// count the number of friends and identify None value and store it as 0 friends_count
String[] friendsList = friends.split(",");
long friends_count = 0;
if (friendsList.length == 1 && friendsList[0].equals("None")) {
    friends_count = 0;
} else {
    friends_count = friendsList.length;
}
```

## b. Data Separation in Yelp Business.json and Yelp User.json

When separating data using commas as separators, issues can arise. For example, when dealing with names that include a comma in between the first and last name, the comma may be mistakenly treated as a separator, leading to incorrect separation of the data. Our solution is that using spaces to replace commas for values in specific fields can clean up the data, help alleviate this issue and ensure accurate separation of the data when creating tables in Hive.



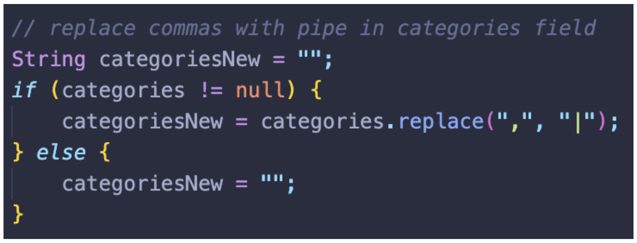
```
String user_id = (String) json.get("user_id");
String name = ((String) json.get("name")).replace(",", ""); // replace commas with space
```



```
String business_id = (String) json.get("business_id");
String name = ((String) json.get("name")).replace(",", ""); // replace commas with space
String city = ((String) json.get("city")).replace(",", ""); // replace commas with space
```

## c. Handling Long and Messy String Data in Yelp Business.json

The Yelp Business.json contains many columns with very long and messy strings, such as the "categories" field. This can make the data difficult to work with and analyze. Our solution is to modify values in the "categories" field by replacing commas with the pipe symbol "|" to avoid values being separated incorrectly when creating tables in Hive. And also, it is easy to read subcategories when analyzing data in Hive and Tableau.



```
// replace commas with pipe in categories field
String categoriesNew = "";
if (categories != null) {
    categoriesNew = categories.replace(",", "|");
} else {
    categoriesNew = "";
}
```

d. **Improve efficiency with Python to clean sample data on local first**

Though we could use the Data Ingest Console to transfer large files to HDFS, and then run MapReduce codes on Hadoop to verify whether the MapReduce jobs work successfully, it is still not time-efficiency.
Our solution is that using Python for data cleaning with sample data on a local is a useful and efficient way to verify that the cleaning process produces correct and valid data for further usage in Hive and Tableau. Python libraries such as JSON and Pandas make tests on codes easier and more efficient. Once the data cleaning process has been validated, MapReduce codes can be written in Java and run on a larger dataset using a Hadoop cluster.

## 4.3. Data Analysis

After data cleaning and profiling, we joined the datasets and analyzed the joined table by using Hive on HDFS. Firstly, we joined the Business, Review and User datasets by using business_id and user_id as foreign keys, then performed analysis to provide insights into the success patterns of popular restaurants in the US and user behavior on Yelp.

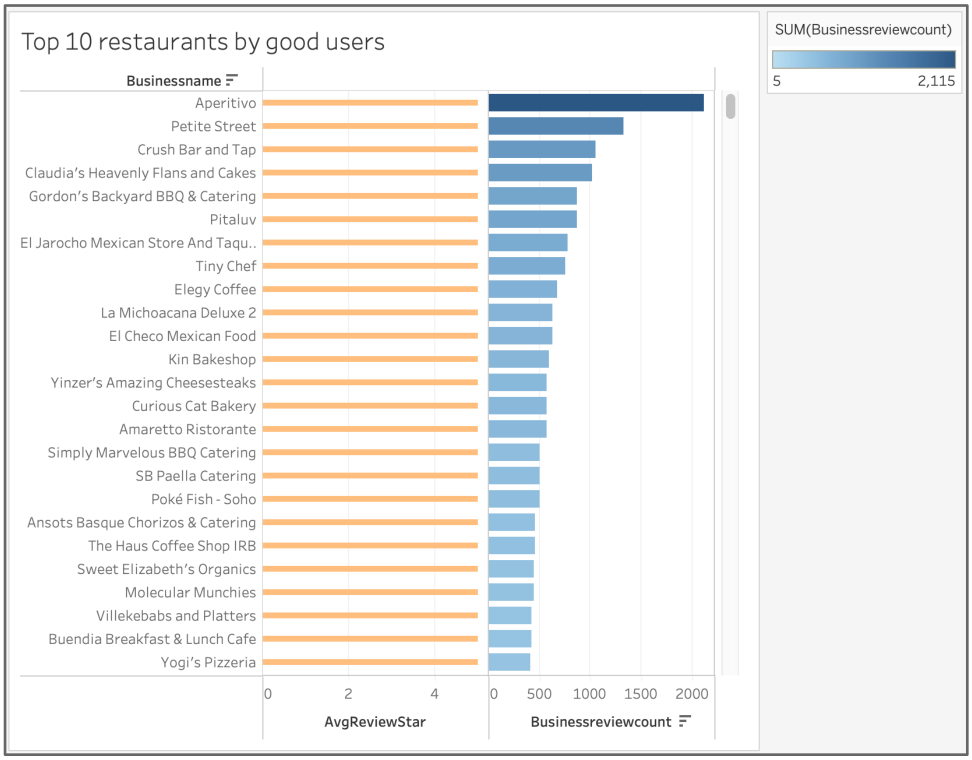
In the analysis of Businesses, we select users with an average star rating higher than 3 as ‘good_users’, and use their ratings to evaluate a business to ensure the quality of reviews. We first select the top 10 restaurants with highest average ranking and review count by good users, and then select top 10 Businesses with Highest Review Variability by 'Good' Users to show the businesses of different categories that have highly polarized reviews; We also selected top 10 businesses with the first category with highest review count and highest average rating to see the most popular and best rating businesses in different business types.

In the analysis of user behavior, we have computed user average review count, average user rating, and Number of Yelp users added per year from the joined table.

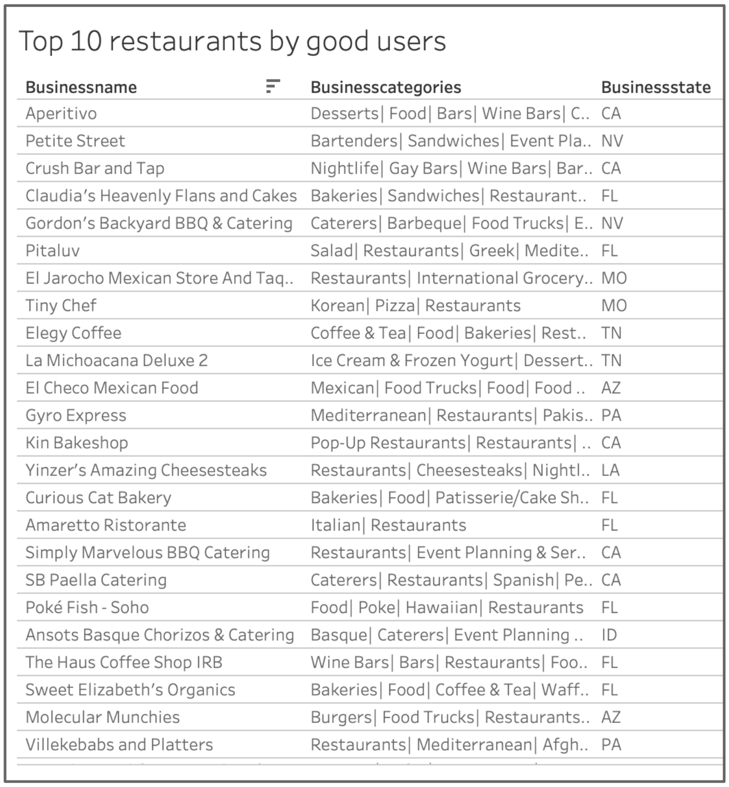
## 4.4. Data Visualization

After analyzing data on Hadoop, we visualized the result by using Tableau to transform the results into compelling and engaging visualizations.

## a. Top 10 restaurants by good users



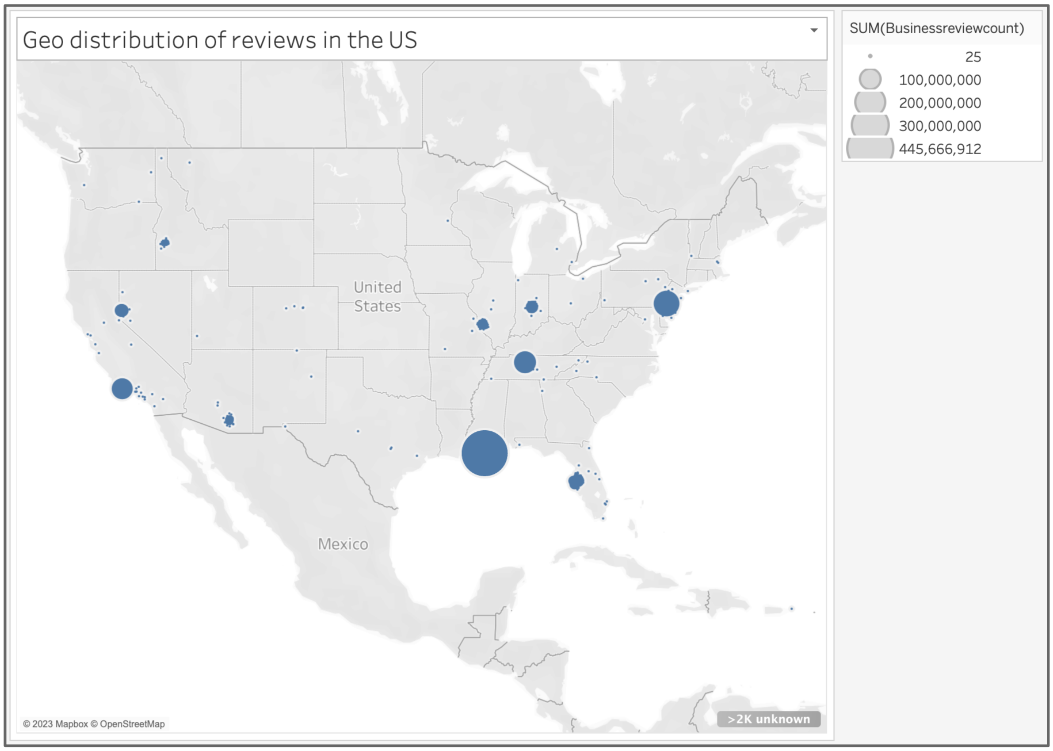
## b. Top 10 restaurants by good users with business categories and state information



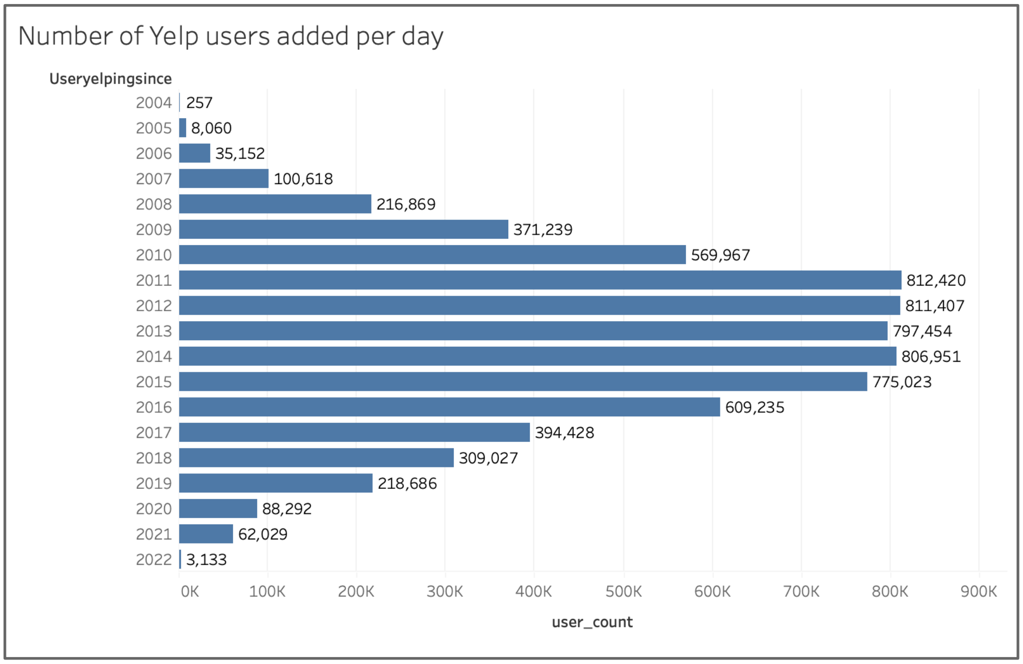
Top 10 restaurants by good users

| Businessname | Businesscategories | Businessstate |
|---|---|---|
| Aperitivo | Desserts\| Food\| Bars\| Wine Bars\| C.. | CA |
| Petite Street | Bartenders\| Sandwiches\| Event Pla.. | NV |
| Crush Bar and Tap | Nightlife\| Gay Bars\| Wine Bars\| Bar.. | CA |
| Claudia's Heavenly Flans and Cakes | Bakeries\| Sandwiches\| Restaurant.. | FL |
| Gordon's Backyard BBQ & Catering | Caterers\| Barbeque\| Food Trucks\| E.. | NV |
| Pitaluv | Salad\| Restaurants\| Greek\| Medite.. | FL |
| El Jarocho Mexican Store And Taq.. | Restaurants\| International Grocery.. | MO |
| Tiny Chef | Korean\| Pizza\| Restaurants | MO |
| Elegy Coffee | Coffee & Tea\| Food\| Bakeries\| Rest.. | TN |
| La Michoacana Deluxe 2 | Ice Cream & Frozen Yogurt\| Dessert.. | TN |
| El Checo Mexican Food | Mexican\| Food Trucks\| Food\| Food .. | AZ |
| Gyro Express | Mediterranean\| Restaurants\| Pakis.. | PA |
| Kin Bakeshop | Pop-Up Restaurants\| Restaurants\| .. | CA |
| Yinzer's Amazing Cheesesteaks | Restaurants\| Cheesesteaks\| Nightl.. | LA |
| Curious Cat Bakery | Bakeries\| Food\| Patisserie/Cake Sh.. | FL |
| Amaretto Ristorante | Italian\| Restaurants | FL |
| Simply Marvelous BBQ Catering | Restaurants\| Event Planning & Ser.. | CA |
| SB Paella Catering | Caterers\| Restaurants\| Spanish\| Pe.. | CA |
| Poké Fish - Soho | Food\| Poke\| Hawaiian\| Restaurants | FL |
| Ansots Basque Chorizos & Catering | Basque\| Caterers\| Event Planning .. | ID |
| The Haus Coffee Shop IRB | Wine Bars\| Bars\| Restaurants\| Foo.. | FL |
| Sweet Elizabeth's Organics | Bakeries\| Food\| Coffee & Tea\| Waff.. | FL |
| Molecular Munchies | Burgers\| Food Trucks\| Restaurants.. | AZ |
| Villekebabs and Platters | Restaurants\| Mediterranean\| Afgh.. | PA |

### c. Geographical distribution of reviews in the U.S.



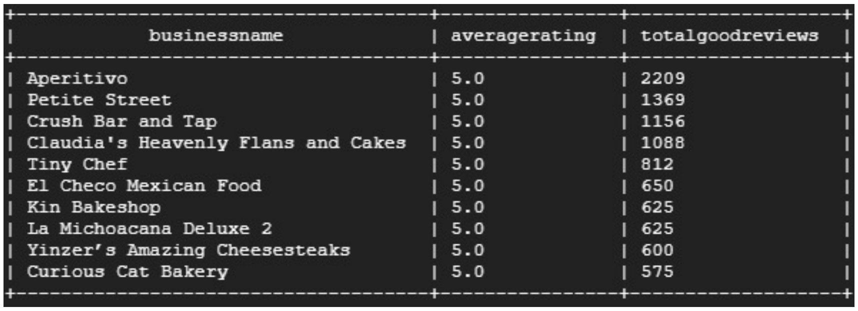
### d. Number of Yelp users added per year



# 5. Results

## a. Top 10 restaurants by good users

Selected top 10 restaurants with highest average ranking and review count by good users whose average rating stars >3 .

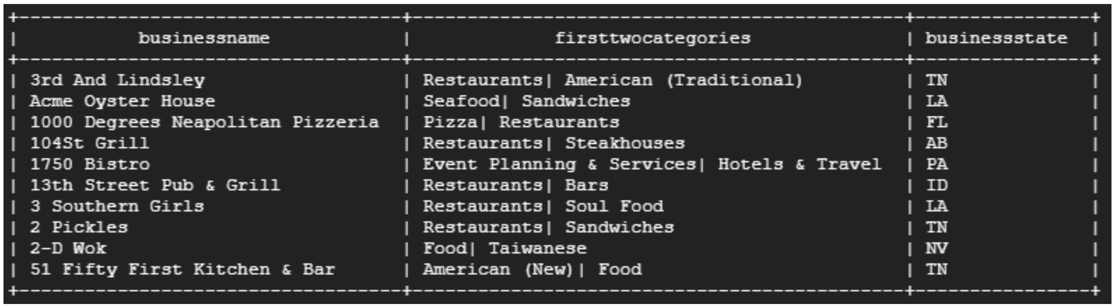


```
+------------------------------------+---------------+------------------+
|            businessname            | averagerating  | totalgoodreviews |
+------------------------------------+---------------+------------------+
| Aperitivo                          | 5.0           | 2209             |
| Petite Street                      | 5.0           | 1369             |
| Crush Bar and Tap                  | 5.0           | 1156             |
| Claudia's Heavenly Flans and Cakes | 5.0           | 1088             |
| Tiny Chef                          | 5.0           | 812              |
| El Checo Mexican Food              | 5.0           | 650              |
| Kin Bakeshop                       | 5.0           | 625              |
| La Michoacana Deluxe 2             | 5.0           | 625              |
| Yinzer’s Amazing Cheesesteaks      | 5.0           | 600              |
| Curious Cat Bakery                 | 5.0           | 575              |
+------------------------------------+---------------+------------------+
```

The highest ranking restaurant is Aperitivo, with 5.0 average rating star and 2209 Reviews.

## b. Top 10 restaurants by good users where yelping_since = 2022 or 2021 or 2020

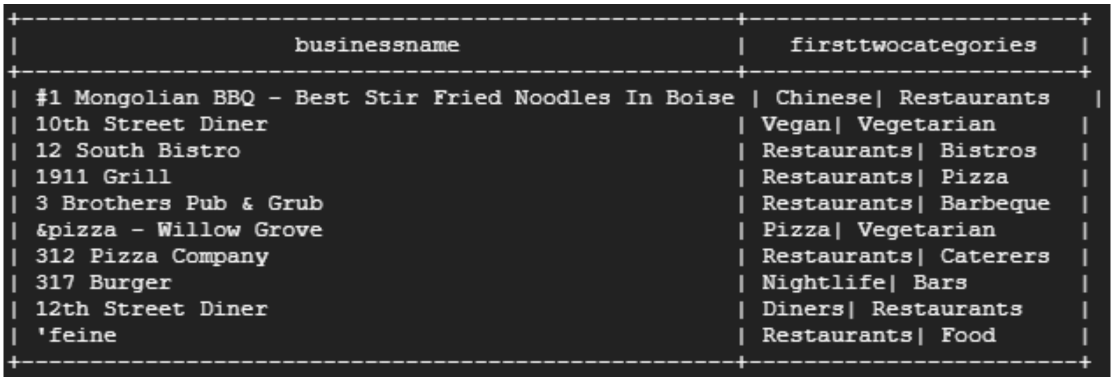
Selected top 10 restaurants with highest average ranking and review count by good users whose average rating stars >3 and joined yelp in recent years.



```
+----------------------------------+--------------------------------------------+----------------+
|           businessname           |              firsttwocategories            | businessstate  |
+----------------------------------+--------------------------------------------+----------------+
| 3rd And Lindsley                 | Restaurants| American (Traditional)        | TN             |
| Acme Oyster House                | Seafood| Sandwiches                        | LA             |
| 1000 Degrees Neapolitan Pizzeria | Pizza| Restaurants                         | FL             |
| 104St Grill                      | Restaurants| Steakhouses                   | AB             |
| 1750 Bistro                      | Event Planning & Services| Hotels & Travel | PA             |
| 13th Street Pub & Grill          | Restaurants| Bars                          | ID             |
| 3 Southern Girls                 | Restaurants| Soul Food                     | LA             |
| 2 Pickles                        | Restaurants| Sandwiches                    | TN             |
| 2-D Wok                          | Food| Taiwanese                            | NV             |
| 51 Fifty First Kitchen & Bar     | American (New)| Food                       | TN             |
+----------------------------------+--------------------------------------------+----------------+
```

c. **Top 10 restaurants reviewed by good users and opinion leaders (who has friends more than avg friends_count)**

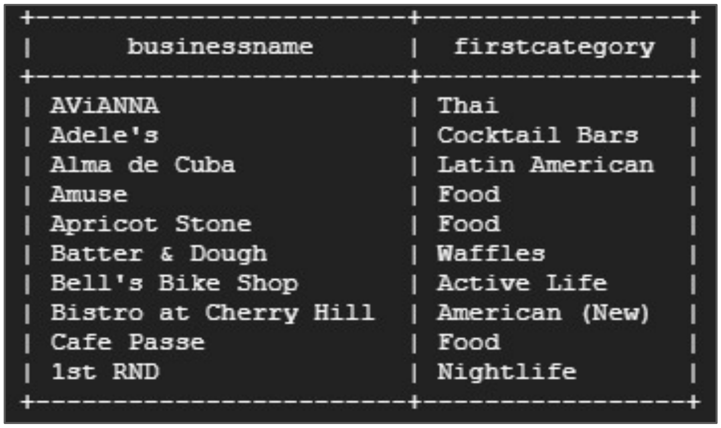
Selected top 10 restaurants with highest average ranking and review count by good users whose average rating stars >3 and who are also opinion leaders in terms of having more friends than average friends_count.



```
+----------------------------------------------------+------------------------+
|                    businessname                    |   firsttwocategories   |
+----------------------------------------------------+------------------------+
| #1 Mongolian BBQ - Best Stir Fried Noodles In Boise | Chinese| Restaurants   |
| 10th Street Diner                                  | Vegan| Vegetarian      |
| 12 South Bistro                                    | Restaurants| Bistros   |
| 1911 Grill                                         | Restaurants| Pizza     |
| 3 Brothers Pub & Grub                              | Restaurants| Barbeque  |
| &pizza - Willow Grove                              | Pizza| Vegetarian      |
| 312 Pizza Company                                  | Restaurants| Caterers  |
| 317 Burger                                         | Nightlife| Bars        |
| 12th Street Diner                                  | Diners| Restaurants    |
| 'feine                                             | Restaurants| Food      |
+----------------------------------------------------+------------------------+
```

d. **Top 10 Businesses with Highest Review Variability by 'Good' Users**

Selected top 10 Businesses with Highest Review Variability by good users whose average rating stars >3.

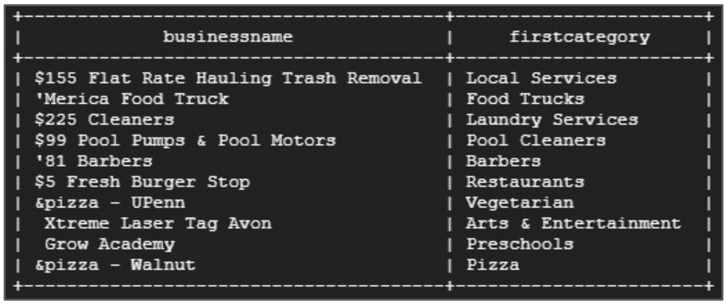


```
+-----------------------+-----------------+
|      businessname     |  firstcategory  |
+-----------------------+-----------------+
| AViANNA               | Thai            |
| Adele's               | Cocktail Bars   |
| Alma de Cuba          | Latin American  |
| Amuse                 | Food            |
| Apricot Stone         | Food            |
| Batter & Dough        | Waffles         |
| Bell's Bike Shop      | Active Life     |
| Bistro at Cherry Hill | American (New)  |
| Cafe Passe            | Food            |
| 1st RND               | Nightlife       |
+-----------------------+-----------------+
```

The highest Review Variability business is AVIANNA, and the business category is Thai.

e. **Top 10 review count business with first categories**

Selected top 10 businesses with the first category with highest review count to see the most popular businesses in different business types.

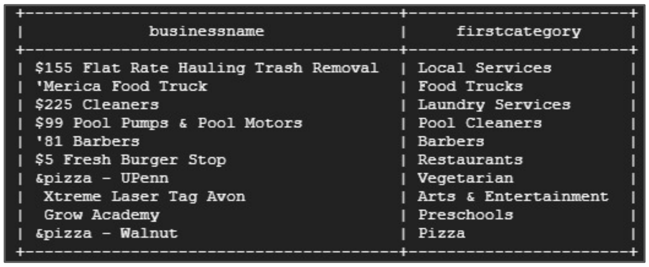


```
+--------------------------------------+-----------------------+
|             businessname             |     firstcategory     |
+--------------------------------------+-----------------------+
| $155 Flat Rate Hauling Trash Removal | Local Services        |
| 'Merica Food Truck                   | Food Trucks           |
| $225 Cleaners                        | Laundry Services      |
| $99 Pool Pumps & Pool Motors         | Pool Cleaners         |
| '81 Barbers                          | Barbers               |
| $5 Fresh Burger Stop                 | Restaurants           |
| &pizza - UPenn                       | Vegetarian            |
|  Xtreme Laser Tag Avon               | Arts & Entertainment  |
|  Grow Academy                        | Preschools            |
| &pizza - Walnut                      | Pizza                 |
+--------------------------------------+-----------------------+
```

The most popular types of business are Local Services, Food Trucks and Laundry Services.

**f. Top 10 rating business with first category**

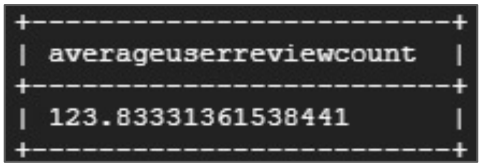
Selected top 10 businesses with the first category with highest average rating to see the best rating businesses in different business types.



```
+--------------------------------------+-----------------------+
|             businessname             |     firstcategory     |
+--------------------------------------+-----------------------+
| $155 Flat Rate Hauling Trash Removal | Local Services        |
| 'Merica Food Truck                   | Food Trucks           |
| $225 Cleaners                        | Laundry Services      |
| $99 Pool Pumps & Pool Motors         | Pool Cleaners         |
| '81 Barbers                          | Barbers               |
| $5 Fresh Burger Stop                 | Restaurants           |
| &pizza - UPenn                       | Vegetarian            |
|  Xtreme Laser Tag Avon               | Arts & Entertainment  |
|  Grow Academy                        | Preschools            |
| &pizza - Walnut                      | Pizza                 |
+--------------------------------------+-----------------------+
```

The most popular businesses with the highest reviews count and the best businesses with the highest average rating are highly identical. The highest rating types of business are also Local Services, Food Trucks and Laundry Services.

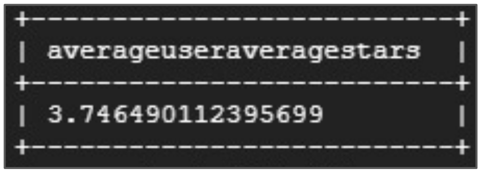
**g. User average review count**



```
+------------------------+
| averageuserreviewcount |
+------------------------+
| 123.83331361538441     |
+------------------------+
```

The average user review count is 123.83.

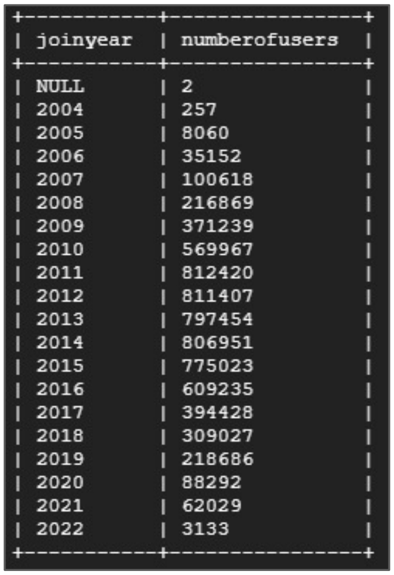
**h. Average user rating**



```
+------------------------+
| averageuseraveragestars |
+------------------------+
| 3.746490112395699      |
+------------------------+
```

The average user rating stars is 3.75.

i. **Number of Yelp users added per day**



```
+-----------+----------------+
| joinyear  | numberofusers  |
+-----------+----------------+
| NULL      | 2              |
| 2004      | 257            |
| 2005      | 8060           |
| 2006      | 35152          |
| 2007      | 100618         |
| 2008      | 216869         |
| 2009      | 371239         |
| 2010      | 569967         |
| 2011      | 812420         |
| 2012      | 811407         |
| 2013      | 797454         |
| 2014      | 806951         |
| 2015      | 775023         |
| 2016      | 609235         |
| 2017      | 394428         |
| 2018      | 309027         |
| 2019      | 218686         |
| 2020      | 88292          |
| 2021      | 62029          |
| 2022      | 3133           |
+-----------+----------------+
```

Yelp's highest number of new users was from 2011 to 2015, after which it showed a year-on-year decline

# 6. Conclusion - Insights

In the analysis of popular restaurants on Yelp, we can see that the most popular restaurants with the highest rating by good users are mostly in state CA, NV, MO, TN, PA and AZ, and different states show their preference on different kinds of cuisine, for example, we observe that in CA, popular restaurants are bars or catering while in PA, popular restaurants are Mediterranean food, and in FL, popular restaurants are foods like salad, bakery and poke. These preferences could have comprehensive analysis with population, lifestyle and weather dataset with corresponding states.

In the analysis of Businesses, we can see that the most popular businesses with the highest reviews count and the best businesses with highest average rating are highly identical. Which means that popularity and service quality are highly positively correlated, so a company can attract more customers by improving the quality of its food and service.

By analyzing the number of reviews and average ratings for all types of business, we found that the top three most popular types of business are Local Services, Food Trucks and Laundry Services, So if there are users who want to start a business, they can consider choosing these areas and will have a greater chance of success.

We also analyzed the variance of business rating. This analysis allowed us to know which companies' scores are highly polarized, and these companies can look deeper into the causes of this phenomenon and make improvements.

In the analysis of user behavior, we found that Yelp's highest number of new users was from 2011 to 2015, after which it showed a year-on-year decline, suggesting that Yelp could address this decline in new users by making improvements, such as developing new features or new incentives to attract more users.

## 7. Acknowledgments

We would like to express our sincere gratitude to HPC for their invaluable support and assistance throughout this project. Their expertise and resources were instrumental in enabling us to effectively analyze and process the large Yelp dataset. We would also like to extend our thanks to Kaggle for providing us with the Yelp dataset that served as the foundation for our analysis. Without the support of HPC and the resources provided by Kaggle, this project would not have been possible. We are truly grateful for their contributions and assistance.

## 8. References

[1]https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset?select=yelp_academic_dataset_business.json
[2]https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset?select=yelp_academic_dataset_user.json
[3]https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset?select=yelp_academic_dataset_review.json
[4]https://data.cityofnewyork.us/City-Government/Demographic-Statistics-By-Zip-Code/kku6-nxdu
[5]https://github.com/yashaz/RBDA/blob/main/RBDA%20Project%20Report%20Spring%202022.pdf