ST 595
Fall 2020
11/8/20

Technical Report: Project 3
Big Data Analysis in Google Cloud Platform

Prepared By:

Josh Olsen

Table of Contents

Introduction

Yelp.com is a database where users can review places like restaurants, resorts, golf courses, theme parks, etc. Users are able check-in at places which can be shared with social media like facebook. I used Yelp to find new restaurants in my neighborhood and to see if they have positive reviews.

Being able to wrangle large amounts of data to provide business insights can provide a significant advantage when trying to improve Key Performance Indicators (K.P.I.).

An analytics manager can determine where to make intelligent marketing investments. For example, if usage on Yelp is decreasing for a certain region, a analytics manager can put an action in place (such as increase marketing spend) to track if the K.P.I.s are increasing due to the increase marketing spend.

Objectives

Below are a series of analytical business questions that will be answered using the Google Cloud Platform.

- What are the top 2 restaurants in Arizona (AZ) that have the most check-ins in June 2016?
  Also, what are their total review counts?

- What State had the most check-ins in January of 2016? And
  What State had the least check-ins in March of 2016?

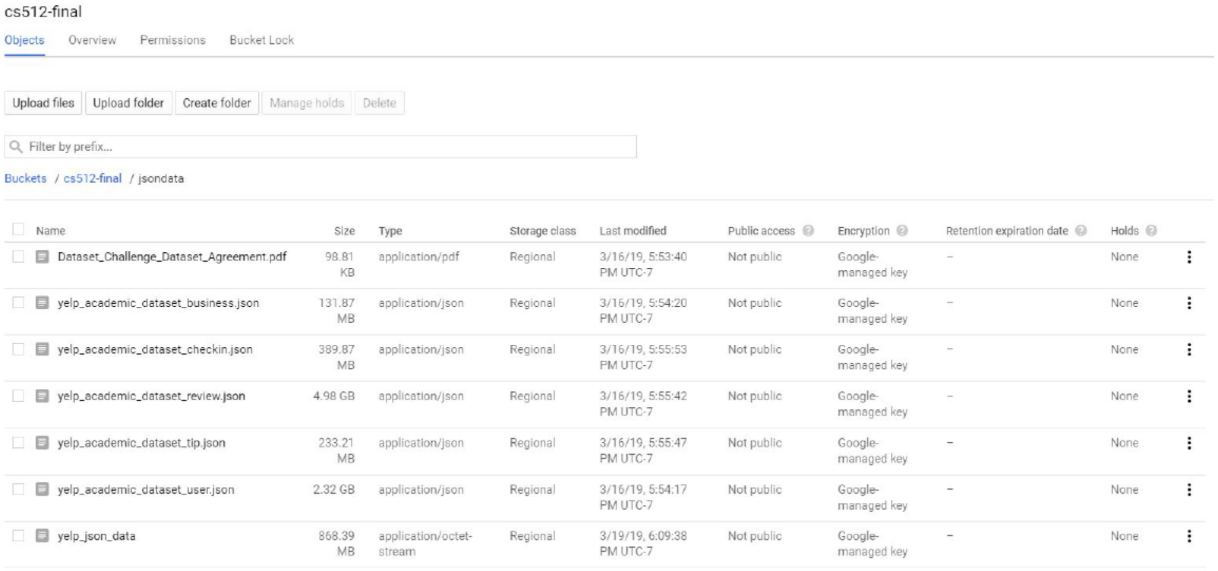
- In the entire dataset, which user has the most reviews? And
  How many reviews does that user have?

Data Description

From my analysis, the Yelp.com dataset has over 6 million user reviews, information on almost 200,000 businesses across 11 cities from January 2004 to December 2018. The dataset also contains check-ins for each user to a business that I thought was very interesting. The check-in data could be used to analyze to show when businesses in your area are trending up or down. Also, the check-in data could be used to incentive users to check-in more so Yelp can get more data to analyze.

Description of dataset

The dataset is composed of five different json files that totals about ten gigabytes of data.

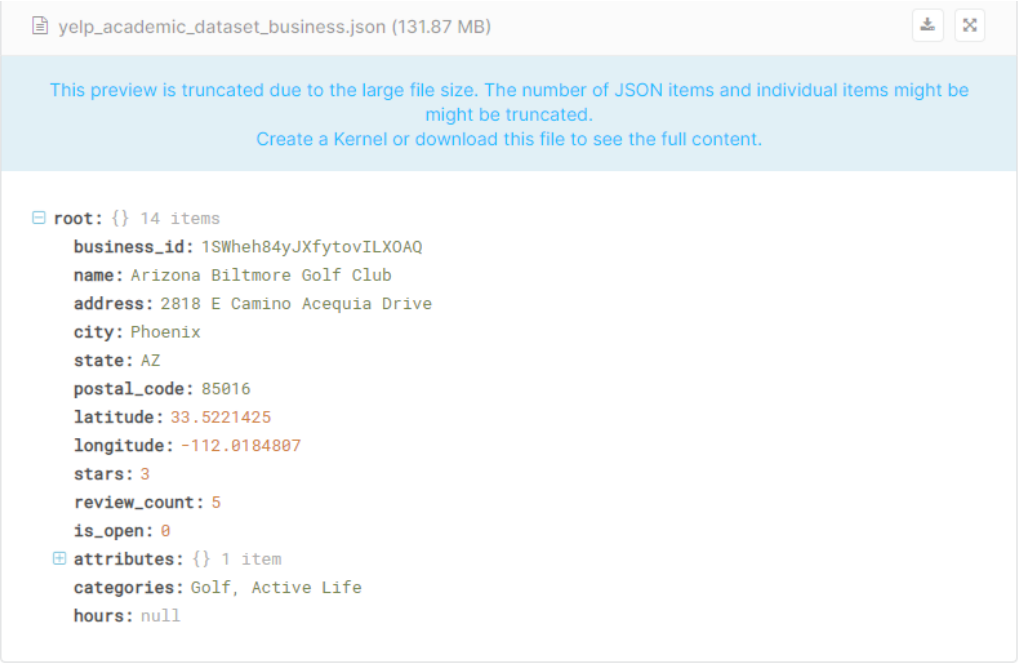


□ Name	Size	Type	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
□ Dataset_Challenge_Dataset_Agreement.pdf	98.81 KB	application/pdf	Regional	3/16/19, 5:53:40 PM UTC-7	Not public	Google-managed key	–	None
□ yelp_academic_dataset_business.json	131.87 MB	application/json	Regional	3/16/19, 5:54:20 PM UTC-7	Not public	Google-managed key	–	None
□ yelp_academic_dataset_checkin.json	389.87 MB	application/json	Regional	3/16/19, 5:55:53 PM UTC-7	Not public	Google-managed key	–	None
□ yelp_academic_dataset_review.json	4.98 GB	application/json	Regional	3/16/19, 5:55:42 PM UTC-7	Not public	Google-managed key	–	None
□ yelp_academic_dataset_tip.json	233.21 MB	application/json	Regional	3/16/19, 5:55:47 PM UTC-7	Not public	Google-managed key	–	None
□ yelp_academic_dataset_user.json	2.32 GB	application/json	Regional	3/16/19, 5:54:17 PM UTC-7	Not public	Google-managed key	–	None
□ yelp_json_data	868.39 MB	application/octet-stream	Regional	3/19/19, 6:09:38 PM UTC-7	Not public	Google-managed key	–	None

Figure 1 - Original Dataset Summary

Figure 1 displays all the Yelp datasets in the Google Cloud Platform environment.
Below in figures 2 - 6 displays examples of each dataset.

yelp_academic_dataset_business.json (131.87 MB)



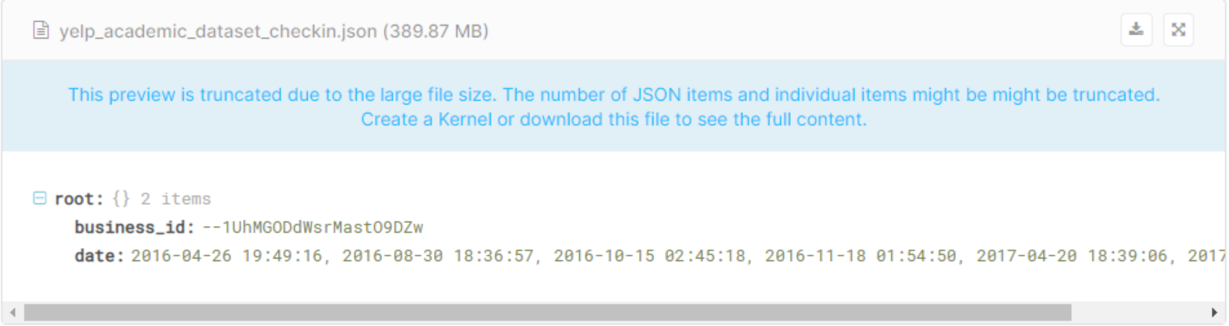
yelp_academic_dataset_business.json (131.87 MB)

This preview is truncated due to the large file size. The number of JSON items and individual items might be might be truncated.
Create a Kernel or download this file to see the full content.

```
⊟ root: {} 14 items
    business_id: 1SWheh84yJXfytovILXOAQ
    name: Arizona Biltmore Golf Club
    address: 2818 E Camino Acequia Drive
    city: Phoenix
    state: AZ
    postal_code: 85016
    latitude: 33.5221425
    longitude: -112.0184807
    stars: 3
    review_count: 5
    is_open: 0
  ⊞ attributes: {} 1 item
    categories: Golf, Active Life
    hours: null
```

Figure 2 - Yelp Business Dataset Example

yelp_academic_dataset_checkin.json (389.87 MB)



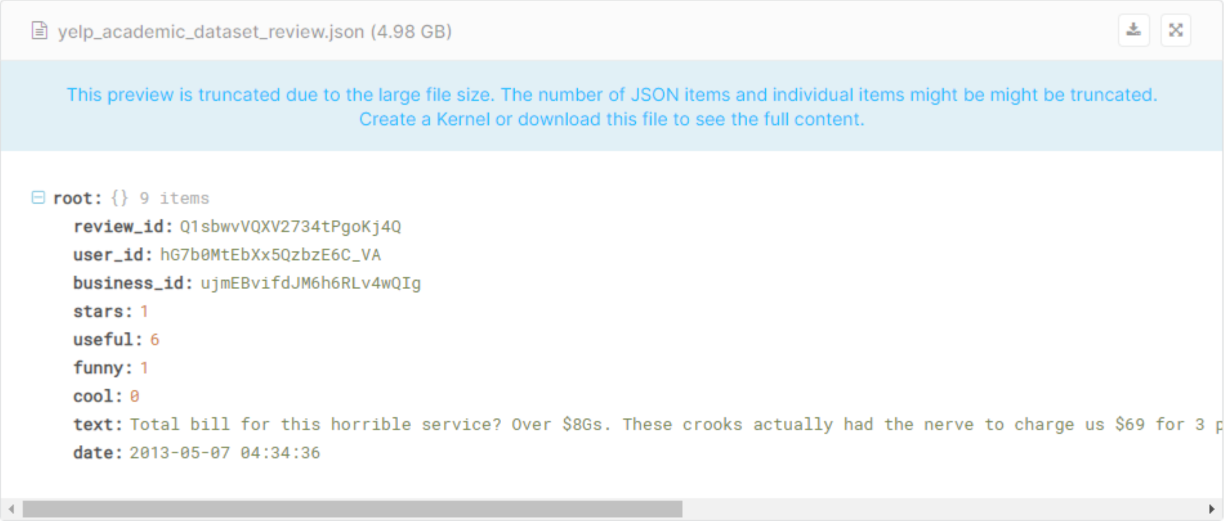
yelp_academic_dataset_checkin.json (389.87 MB)

This preview is truncated due to the large file size. The number of JSON items and individual items might be might be truncated.
Create a Kernel or download this file to see the full content.

```
⊟ root: {} 2 items
    business_id: --1UhMGODdWsrMast09DZw
    date: 2016-04-26 19:49:16, 2016-08-30 18:36:57, 2016-10-15 02:45:18, 2016-11-18 01:54:50, 2017-04-20 18:39:06, 2017
```

Figure 3 - Yelp Check-In Dataset Example

yelp_academic_dataset_review.json (4.98 GB)



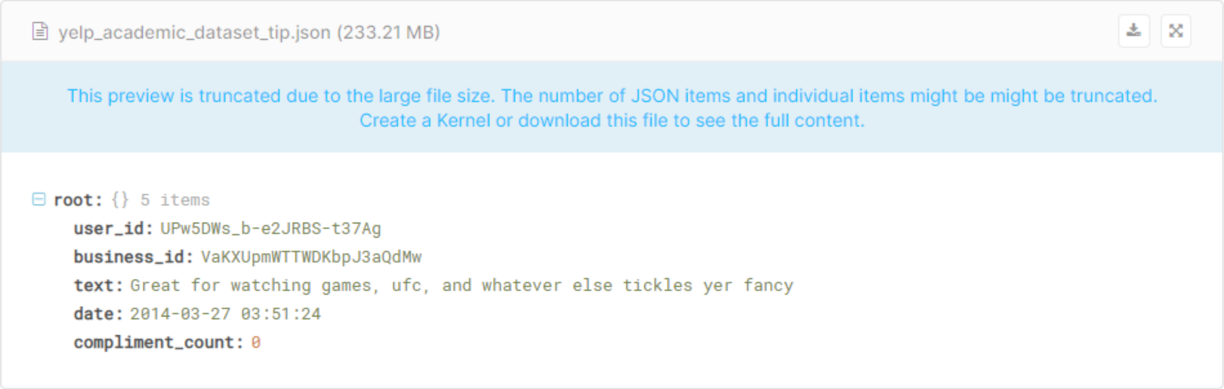
yelp_academic_dataset_review.json (4.98 GB)

This preview is truncated due to the large file size. The number of JSON items and individual items might be might be truncated.
Create a Kernel or download this file to see the full content.

```
⊟ root: {} 9 items
    review_id: Q1sbwvVQXV2734tPgoKj4Q
    user_id: hG7b0MtEbXx5QzbzE6C_VA
    business_id: ujmEBvifdJM6h6RLv4wQIg
    stars: 1
    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 p
    date: 2013-05-07 04:34:36
```

Figure 4 - Yelp Business Reviews Dataset Example

yelp_academic_dataset_tip.json (233.21 MB)



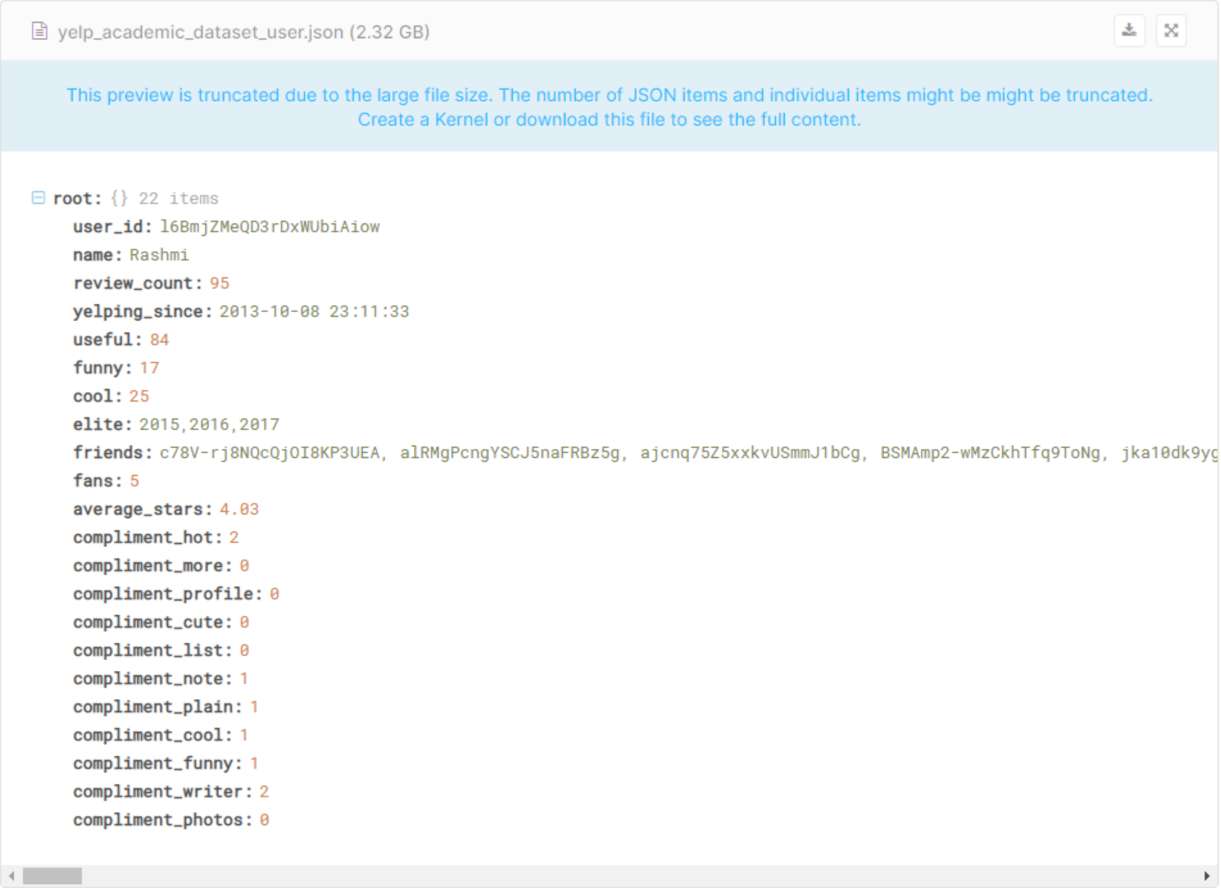
yelp_academic_dataset_tip.json (233.21 MB)

This preview is truncated due to the large file size. The number of JSON items and individual items might be might be truncated.
Create a Kernel or download this file to see the full content.

```
⊟ root: {} 5 items
    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
```

Figure 5 - Yelp Business Tips Dataset Example

yelp_academic_dataset_user.json (2.32 GB)



yelp_academic_dataset_user.json (2.32 GB)

This preview is truncated due to the large file size. The number of JSON items and individual items might be might be truncated.
Create a Kernel or download this file to see the full content.

```
⊟ root: {} 22 items
    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, ajcnq75Z5xxkvUSmmJ1bCg, BSMAmp2-wMzCkhTfq9ToNg, jka10dk9yg
    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
```

Figure 6 - Yelp User Dataset Example

Business Analytics Overview

The Yelp datasets needed to get into the Google Cloud Platform.
Below in Figure 7 visually displays all the datasets directed to the Google Cloud Platform.

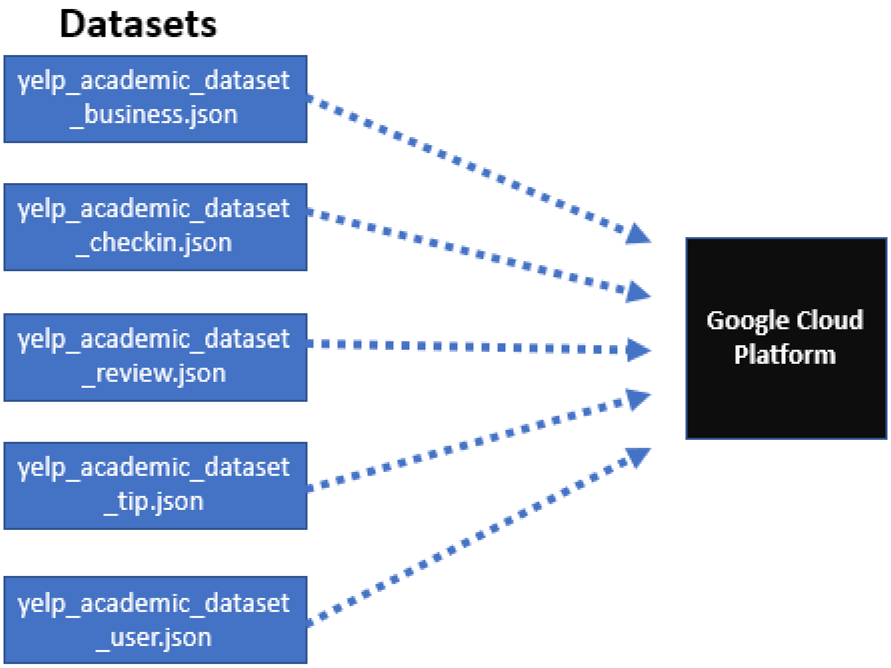

Datasets
yelp_academic_dataset_business.json
yelp_academic_dataset_checkin.json
yelp_academic_dataset_review.json
yelp_academic_dataset_tip.json
yelp_academic_dataset_user.json
Google Cloud Platform

Figure 7 - Datasets and Google Cloud Platform Overview

There are key steps to be able to answer the analytical questions. After the datasets are in the Google Cloud Storage, the data was prepared in DataPrep to be able to Query in BigQuery.

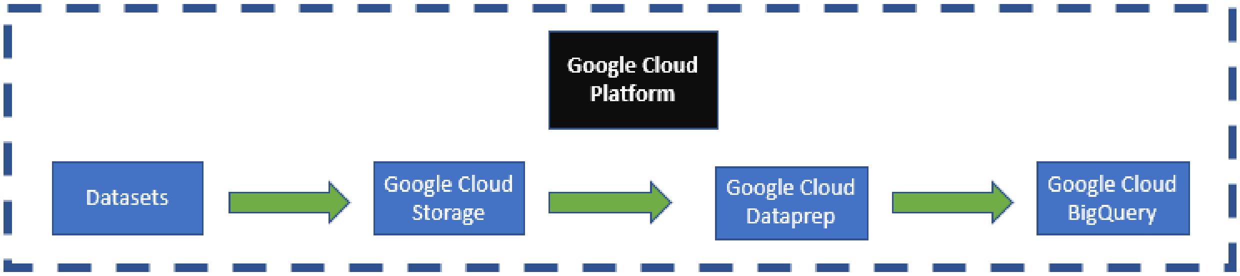


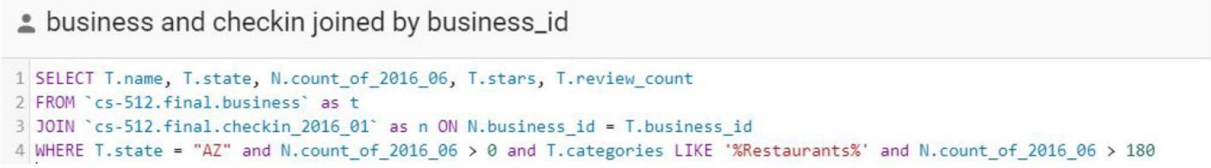
Figure 8 - Google Cloud Platform Environment

Analytical Questions:

- What are the top 2 restaurants in Arizona (AZ) that have the most check-ins in June 2016?
  Also, what are their total review counts?

The business check-in table had the count of check-ins for June of 2016.
To be able to answer the question, I needed to join the check-in table to the business table so I can query against the count of check-ins to the business.

I filtered the query to only show the state of “AZ”.
Next, I filtered the categories to only include “Restaurants”.
My last filter was to only show the counts that are above 180.
Below is the query I used to get the answer below.



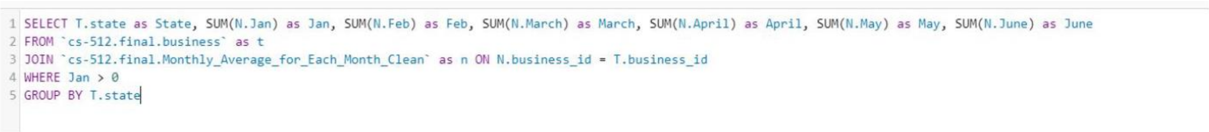
business and checkin joined by business_id

```
SELECT T.name, T.state, N.count_of_2016_06, T.stars, T.review_count
FROM `cs-512.final.business` as t
JOIN `cs-512.final.checkin_2016_01` as n ON N.business_id = T.business_id
WHERE T.state = "AZ" and N.count_of_2016_06 > 0 and T.categories LIKE '%Restaurants%' and N.count_of_2016_06 > 180
```

Figure 9 - BigQuery for Analytical Question 1

- What State had the most check-ins in January of 2016? And
  What State had the least check-ins in March of 2016?

This question required me to join two tables and to be able to sum up the check-ins for each month.
I joined the business table and the Monthly Check-in Count table. This would allow me to get the business info with the check-in info.
To get an aggregate of the each month check-ins, I had to sum all states for Jan - June of 2016.
See SELECT section in the snip below.
Next, I joined the business table to the Monthly check-in count.
Then I grouped by state.

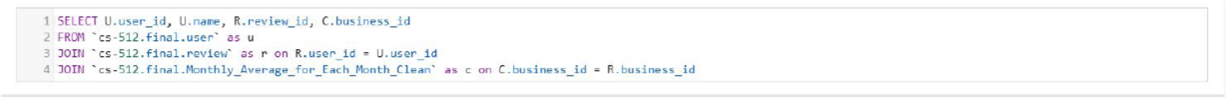


```
SELECT T.state as State, SUM(N.Jan) as Jan, SUM(N.Feb) as Feb, SUM(N.March) as March, SUM(N.April) as April, SUM(N.May) as May, SUM(N.June) as June
FROM `cs-512.final.business` as t
JOIN `cs-512.final.Monthly_Average_for_Each_Month_Clean` as n ON N.business_id = T.business_id
WHERE Jan > 0
GROUP BY T.state
```

Figure 10 - BigQuery for Analytical Question 2

- In the entire dataset, which user has the most reviews? And
  How many reviews does that user have?

This question was probably the hardest to answer because it involved two major steps. Creating a new table from from the user, review, and Monthly Check-in Table get to a master list of reviews.

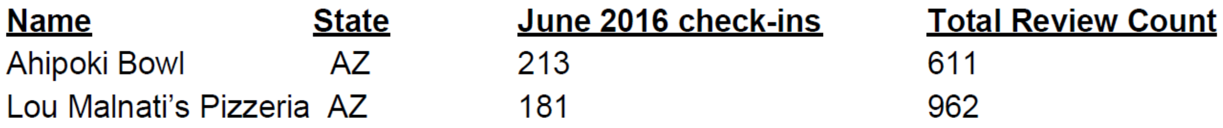


```
SELECT U.user_id, U.name, R.review_id, C.business_id
FROM `cs-512.final.user` as u
JOIN `cs-512.final.review` as r on R.user_id = U.user_id
JOIN `cs-512.final.Monthly_Average_for_Each_Month_Clean` as c on C.business_id = R.business_id
```

Figure 11 - BigQuery for Analytical Question 3

Results

- What are the top 2 restaurants in Arizona (AZ) that have the most check-ins in June 2016?
  Also, what are their total review counts?

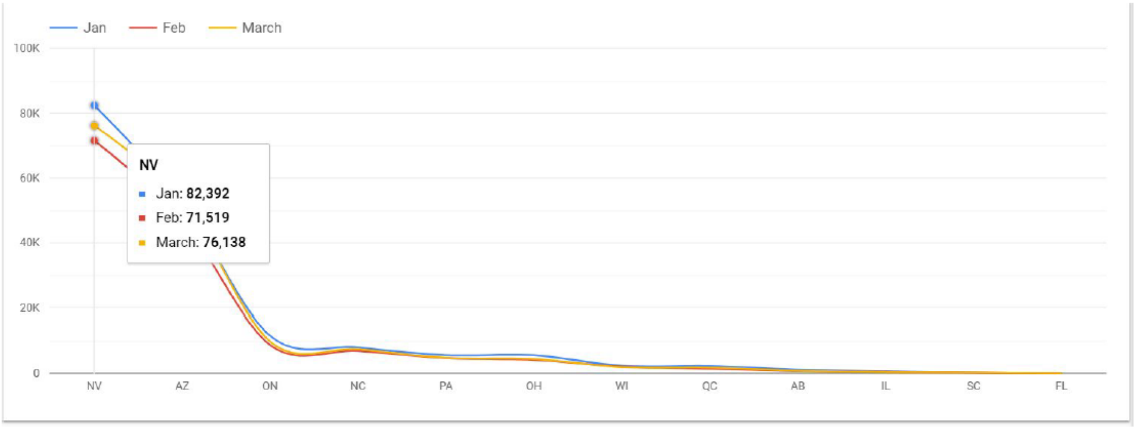


Name	State	June 2016 check-ins	Total Review Count
Ahipoki Bowl	AZ	213	611
Lou Malnati’s Pizzeria	AZ	181	962

Figure 12 - Results for Analytical Question 1

- What State had the most check-ins in January of 2016? And What State had the least check-ins in March of 2016?

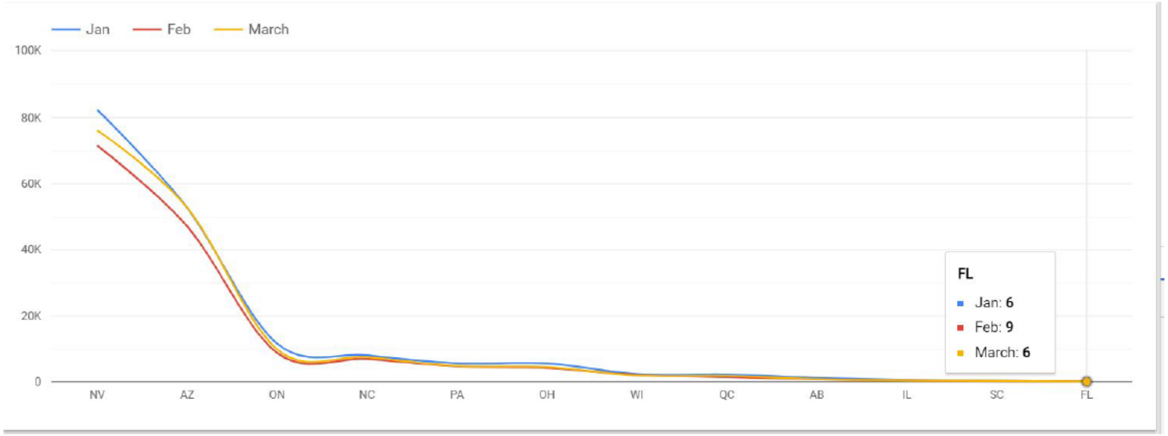
Nevada (NV) had the most amount of check-ins in January of 2016 at 82,392.



Monthly Check-Ins for Each State

Figure 12 - Results for Analytical Question 2 Part 1

Florida (FL) had the least amount of check-ins in March of 2016 at 6.

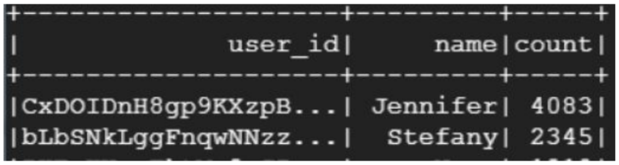


Monthly Check-Ins for Each State

Figure 13 - Results for Analytical Question 2 Part 2

- In the entire dataset, which user has the most reviews? And
  How many reviews does that user have?

Jennifer with user_id of CxDOIDnH8gp9KXzpB had the most check-ins at 4,083.



```
+--------------------+---------+-----+
|             user_id|     name|count|
+--------------------+---------+-----+
|CxDOIDnH8gp9KXzpB...| Jennifer| 4083|
|bLbSNkLggFnqwNNzz...|  Stefany| 2345|
```

Stefany had the second amount of check-ins at 2,345.
Jennifer had 1,738 more check-ins than Stefany.

Figure 13 - Results for Analytical Question 3

Conclusion

Using insights from Big Data can help improve performance. A analytics manager from Yelp could provide incentives to restaurants who have the most check-ins. On the other side, there could be incentives to improve performance for restaurants who do not have a lot of check-ins or who are downtrending.

Another example an analytics manager from Yelp could determine which states are increasing or decreasing with restaurant check-ins. This is a valuable insight to help determine where to increase marketing spend for a certain state that is not performing well against other states.

Yelp user feedback is also a valuable insight. Certain users who reach milestones could be rewarded with incentives to continue to provide data into the Yelp environment. An analytics manager at Yelp could track new users to ensure they are using the Yelp environment.

Future Steps in this Analysis

With this large dataset from Yelp, there are a few future steps to improve this analysis.

First is to help identify risks with restaurants who are about to go out of business with Natural Language Processing & Sentiment Analysis. For example, if a business is constantly getting negative reviews, Yelp could help consult with the business to help fix their problems before they go out of business.

Another future step is using an Advanced Regression Machine Learning Model to determine which features of a popular restaurant tend to have. Such as pricing of much, selection of menu, and neighborhood.

Also, I plan on getting my Google Cloud Platform environment setup to connect to Jupiter notebook to run python.
This will allow me to expand the technical report to include some data science elements such has

- Exploratory Data Analysis,
- Wordcloud, and
- Sentiment of reviews