

**Name: MohammedSaif Shaikh**

**CS 512**

**March 20<sup>th</sup>, 2023**

## **FINAL PROJECT**

**Problem statement:** - The problem of the decline in the growth of businesses in the United States, has an impact on the country's economy, which affects the entire nation, so a good starting point would be to find which States contribute more to business and which ones less & compare the factors affecting the business such as hours, ratings and is currently open.

### **OSEMN PROCESS:** -

#### **Obtain:** -

This data is collected from the yelp dataset which is one of the datasets given in the assignment. This dataset contains information about the business name, states, hours, categories, address, and star rating and is open. This data is used to monitor the overall business situation of the entire nation.

#### **Scrub:** -

Before processing these data, we need to scrub this data. It was a difficult part as the dataset was too large (over 4 GB) to be able to load into the system, along with that it was unstructured. So, firstly I divided it into smaller segments and after dividing it I edited all the sub-dataset in the standardized format (JSON). For this, I used Notepad++ as the files were unable to get opened in visual studio due to memory limitations. After that, I created a bucket in the google cloud storage and uploaded the JSON files that I edited. Now in order to clean these data we need to import that data first and for that, I used data prep where I imported this data which was uploaded to google cloud storage. In data prep, firstly I selected the bucket that contains all the required files of the dataset for importing it was giving an error when I tried to

import files of large size, so I had to create a smaller sized file and imported them which worked for me. Furthermore, in order to clean these data, I created new recipes to handle null values, outliers as well as any mismatched values that may hinder our results. After cleaning the dataset, I sent it to a big query for the analysis process.

### **Explore: -**

In this stage, we will explore this data using GCP. After successfully sending the cleaned datasets from dataprep to big query, I performed analysis on the business and review table that we loaded for 3 questions which 2 were done using big query with looker studio visualization and one using pyspark analysis.

### **Model: -**

To model our data, we can use clustering algorithms for making a cluster of states, classification algorithms for classifying into a group of states, or regression algorithms as per the requirement. In this step, we need to reduce the dimensionality of our data and must select only those data from which we can easily predict the results. This is not required for our dataset as our assignment requirement is limited to loading and analyzing the data.

### **Interpreting: -**

This is the final stage as well as the most problematic stage. In this stage, we are visualizing the data with the bar chart and the results show the top 10 most reviewed restaurants in Philadelphia with their ratings given by their customers. Along with that, we can find the top 10 cities that are having the highest number of businesses and what are the cities with the highest and lowest average ratings among these.

## **OVERVIEW OF DATASET: -**

The Yelp dataset is a subset of businesses, reviews, and user data as JSON files and is a 6-point dataset since it is split up across multiple files, larger than 1 GB, data contains strings with punctuation, a dataset is composed of more than one type of related data. It has 5 different tables which include business, review, check-in, tip, and user. This dataset is a challenging one as the data that we have is not in the standardized format.

## **SAMPLE OF INITIAL DATA: -**

### **Business data:**

```
{
  "business_id": "Pns2l4eNsfO8kk83dixA6A",
  "name": "Abby Rappoport, LAC, CMQ",
  "address": "1616 Chapala St, Ste 2",
  "city": "Santa Barbara",
  "state": "CA",
  "postal_code": "93101",
  "latitude": 34.4266787,
  "longitude": -119.7111968,
  "stars": 5.0,
  "review_count": 7,
  "is_open": 0,
  "attributes": { "ByAppointmentOnly": "True" },
  "categories": "Doctors, Traditional Chinese Medicine, Naturopathic/Holistic,
  Acupuncture, Health
  & Medical, Nutritionists",
  "hours": null
}
....
```

### **Review data:**

```
{
  "review_id": "KU_O5udG6zpxOg-VcAEodg",
  "user_id": "mh_-eMZ6K5RLWhZyISBhwA",
  "business_id": "XQfwVwDr-v0ZS3_CbbE5Xw",
  "stars": 3.0,
  "useful": 0, "funny": 0,
  "cool": 0,
```

```
    "text": "If you decide to eat here, just be aware it is going to take about 2 hours from  
beginning to end. We have tried it multiple times, because I want to like it! I have been to  
it's other locations in NJ and never had a bad experience. \n\nThe food is good, but it  
takes  
a very long time to come out. The waitstaff is very young, but usually pleasant. We have  
just had too many experiences where we spent way too long waiting. We usually opt for  
another diner or restaurant on the weekends, in order to be done quicker.",  
    "date": "2018-07-07 22:09:11"  
}  
....
```

### Check-in data:

```
{  
    "business_id": "---kPU91CF4Lq2-WIRu9Lw",  
    "date": "2020-03-13 21:10:56, 2020-06-02 22:18:06, 2020-07-24 22:42:27, 2020-10-24  
21:36:13, 2020-12-09 21:23:33, 2021-01-20 17:34:57, 2021-04-30 21:02:03, 2021-05-25  
21:16:54, 2021-08-06 21:08:08, 2021-10-02 15:15:42, 2021-11-11 16:23:50"  
}  
.....
```

### Tip data:

```
{  
    "user_id": "AGNUgVwnZUey3gcPCJ76iw",  
    "business_id": "3uLgwr0qeCNMjKenHJwPGQ",  
    "text": "Avengers time with the ladies.",  
    "date": "2012-05-18 02:17:21",  
    "compliment_count": 0  
}  
....
```

### User data:

```
{  
    "user_id": "fJZO_skqpnhk1kvomI4dmA",  
    "name": "Jennifer",  
    "review_count": 25,  
    "yelping_since": "2008-07-14 16:01:36",  
    "useful": 29,  
    "funny": 2,  
    "cool": 19,  
    "elite": "",  
    "friends": "hJiJzw6obCmbGAfwrTkavQ, EMJV9rib660I4RpMsbzWbg,
```

GJv1yf\_IhUZqpDjFr86DmA, h2EmAN1svEbwJqh3H2L7kg,  
ll63altLtfOgVhEM0KITqA",

"fans":1,

"average\_stars":4.15,

"compliment\_hot":0,

"compliment\_more":1,

"compliment\_profile":0,

"compliment\_cute":0,

"compliment\_list":0,

"compliment\_note":6,

"compliment\_plain":2,

"compliment\_cool":2,

"compliment\_funny":2,

"compliment\_writer":1,

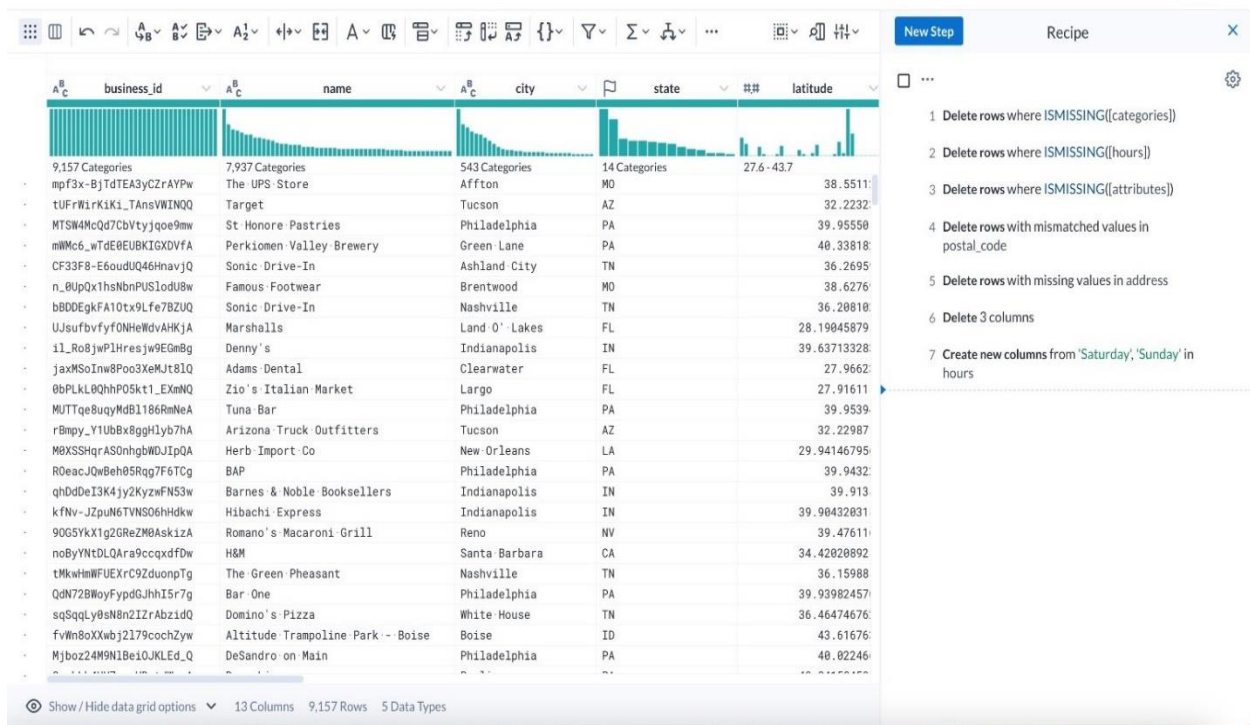
"compliment\_photos":0

}

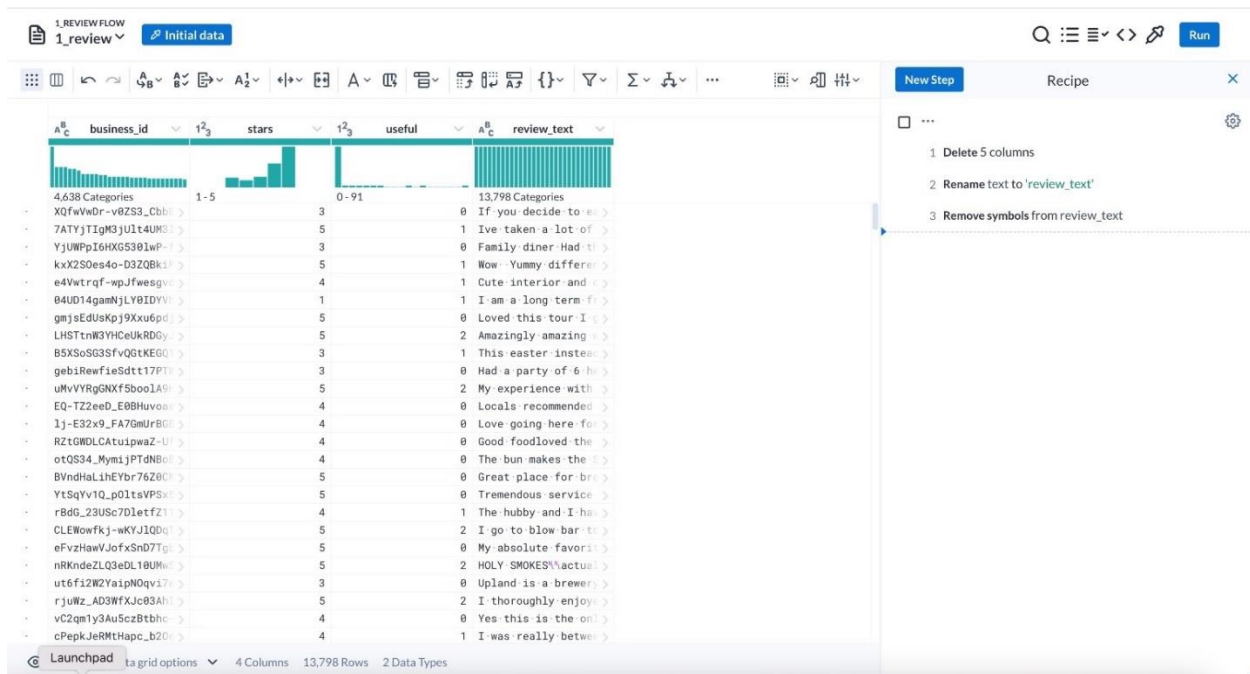
.....

## OVERVIEW OF DATA WRANGLING PROCESS: -

As the initial data were in unstructured form as well as it was too large to be loaded into the local system so I first created a sub-dataset from the original dataset which will decrease the file size then edited those through Notepad++. For cleaning the data by removing errors from noisy data, I used data prep. First, I imported the files that were uploaded to the bucket and then created a new flow in order to apply recipes. I had to use sub-datasets files that I created in order to successfully import it into data prep as it was throwing an error while I import a file size of more than 150 Mb. For my analysis, I needed only business and review files.



For Business files, recipes I created is deleting rows from the categories, hours, attributes, and postal code where there is a null value as these columns are important for my analysis question, then I deleted attributes, address, and postal code column as it was not required.

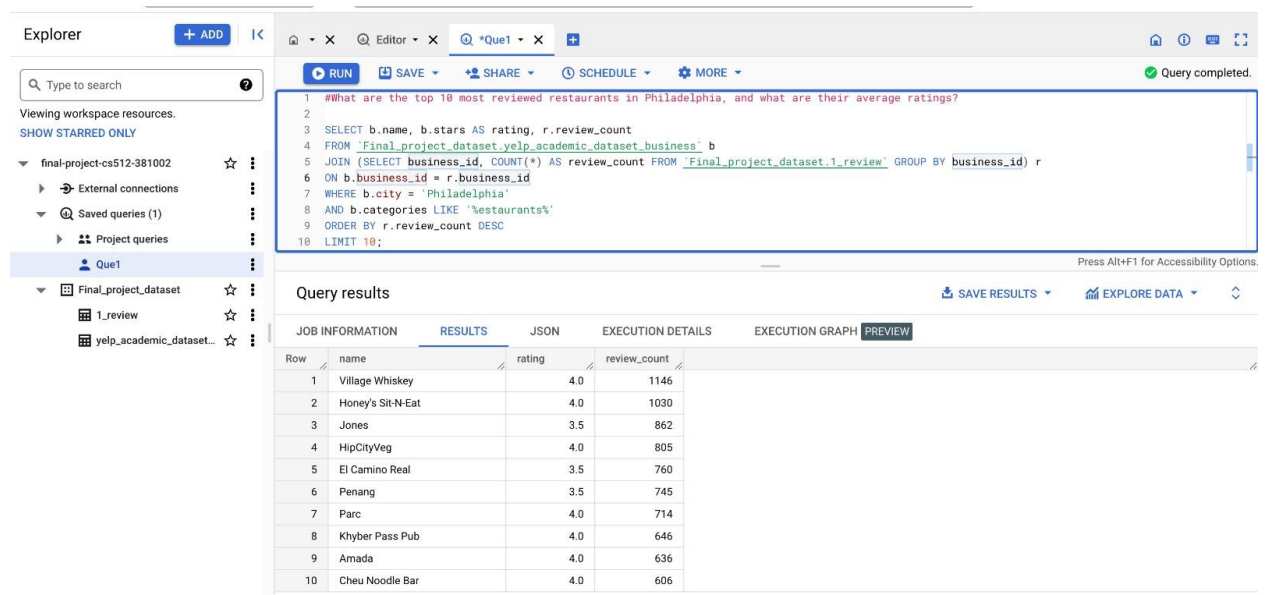


For review, I removed the unnecessary columns that were not necessary. Also, I removed the symbols and converted them to strings having only alpha numeric characters.

## DATA ANALYSIS: -

### 1. What are the top 10 most reviewed restaurants in Philadelphia and what are their ratings (Big Query with visualization)?

To answer this question, first, we need to consider two tables of business and review. From the business table, we need the name of the business and stars, through which we can find the top 10 businesses that are in Philadelphia. Then we need to consider the review table in order to get the review count of those restaurants. For this, we need to join both the business and review tables on business id and then we will sort the data by review count to identify the top 10 most reviewed restaurants.



The screenshot displays the Google BigQuery web interface. On the left, the 'Explorer' pane shows the project 'final-project-cs512-381002' and the dataset 'Final\_project\_dataset' with tables '1\_review' and 'yelp\_academic\_dataset...'. The main editor shows a SQL query to find the top 10 most reviewed restaurants in Philadelphia. The query results are displayed in a table with columns 'name', 'rating', and 'review\_count'.

```
1 #What are the top 10 most reviewed restaurants in Philadelphia, and what are their average ratings?
2
3 SELECT b.name, b.stars AS rating, r.review_count
4 FROM `Final_project_dataset.yelp_academic_dataset.business` b
5 JOIN (SELECT business_id, COUNT(*) AS review_count FROM `Final_project_dataset.1_review` GROUP BY business_id) r
6 ON b.business_id = r.business_id
7 WHERE b.city = 'Philadelphia'
8 AND b.categories LIKE '%restaurants%'
9 ORDER BY r.review_count DESC
10 LIMIT 10;
```

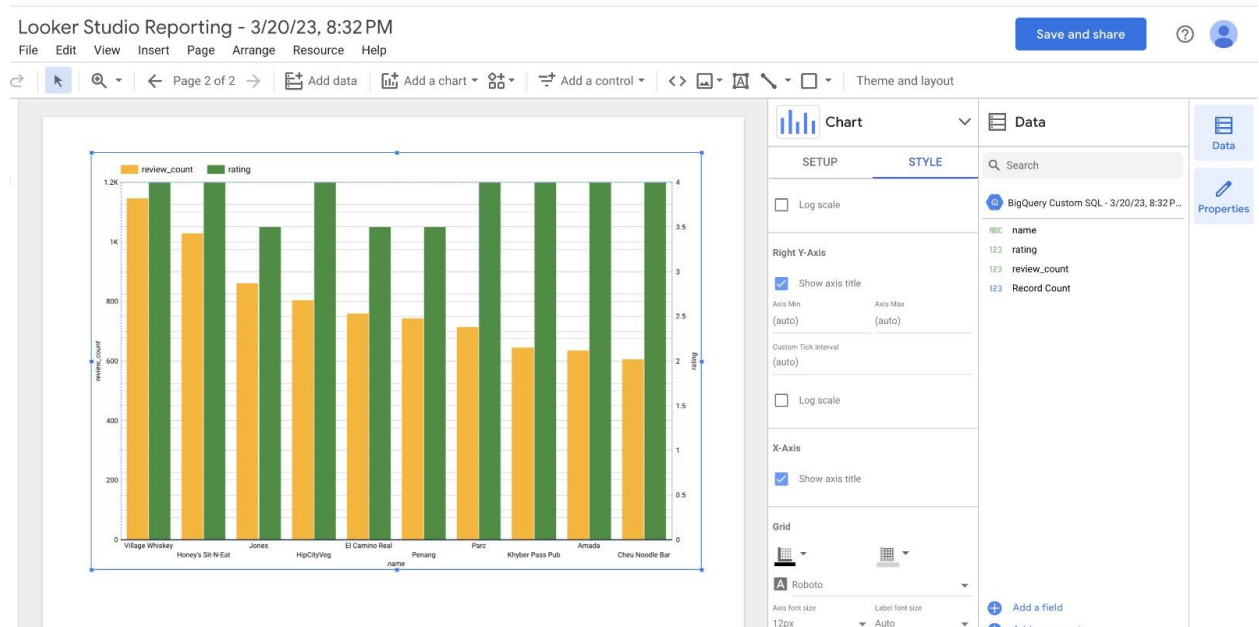
Row	name	rating	review_count
1	Village Whiskey	4.0	1146
2	Honey's Sit-N-Eat	4.0	1030
3	Jones	3.5	862
4	HipCityVeg	4.0	805
5	El Camino Real	3.5	760
6	Penang	3.5	745
7	Parc	4.0	714
8	Khyber Pass Pub	4.0	646
9	Amada	4.0	636
10	Cheu Noodle Bar	4.0	606

### Final Result:

The above analysis provides information about the most popular restaurants in Philadelphia that are having high reviews from their customers. This information can be utilized by businesses to evaluate their competition and assist discover areas for improvement as well as to identify the well-known and highly rated eateries in that specific area.



## Console Output:



## 2. What are the no. of businesses that are opened on Sunday and are highly rated, Statewise? (Pyspark Analysis)

The question aims at the calculating total number of businesses according to the states that are open on Sundays and have high ratings, sorted by states. For this, I gather data on businesses in each state along with their operating hours and ratings. I then filter those businesses to only include those with excellent reviews that are open on Sundays.

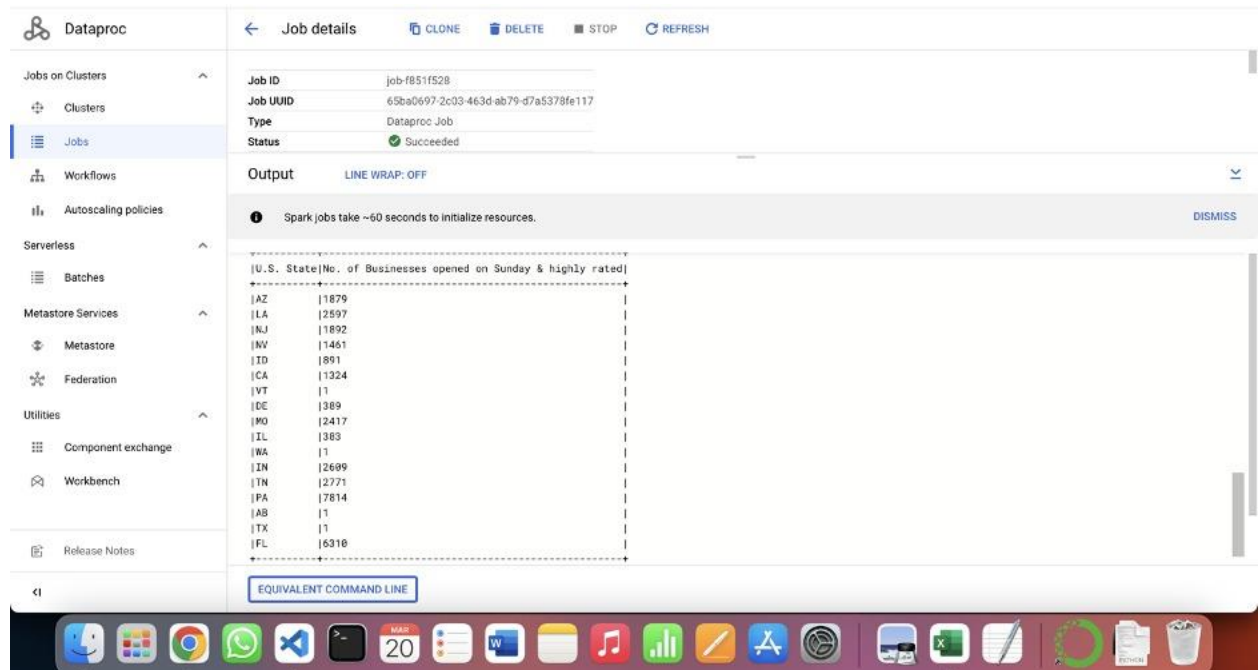
After filtering the data by state and counting the number of Sunday businesses in each state that meets high rating criteria. Then I was able to get the state-by-state breakdown of the number of highly rated Sunday businesses.

## Final Result:

The above analysis provides a state-by-state breakdown of the number of businesses that are open on Sundays and have high ratings. The result from

this analysis will be useful for those trying to locate places with a large concentration of reputable Sunday enterprises or policymakers interested in comprehending trends of trade and consumer behavior.

### Console Output:



The screenshot shows the Databricks console interface. On the left is a sidebar with navigation options: Jobs on Clusters, Clusters, Jobs (selected), Workflows, Autoscaling policies, Serverless, Batches, Metastore Services, Metastore, Federation, Utilities, Component exchange, Workbench, and Release Notes. The main panel displays 'Job details' for a job named 'Dataproc'. The job status is 'Succeeded'. Below the job details, the 'Output' section shows a table with the following data:

[U.S. State]	No. of Businesses opened on Sunday & highly rated
AZ	11879
LA	12697
NJ	11892
NV	11461
TD	891
CA	11324
VT	1
DE	1389
MO	12417
IL	1383
WA	1
IN	12609
TN	12771
PA	17814
AB	1
TX	1
FL	16316

Below the table, there is a button labeled 'EQUIVALENT COMMAND LINE'. The bottom of the screenshot shows a macOS dock with various application icons.

### 3. What are the top 10 cities with the highest no. of businesses and what are the cities with the highest and lowest average ratings among these businesses?

To answer this question first we gather data on businesses in each city, including information on their locations, types, and ratings. Then, we calculate the total number of businesses in each city and rank the cities according to their metric to identify the top 10. After that, I calculate the average ratings for businesses in each city and rank the cities from the highest to the lowest which will help us to identify the cities where businesses tend to have the highest and lowest ratings, respectively.

## Final Result:

The above analysis aimed to identify the top 10 cities with the highest number of businesses and the cities with the highest and lowest ratings among these businesses. These cities are anticipated to be significant economic hubs in their respective regions because of the large concentration of enterprises they contain. Overall, this study can be helpful for a number of things, including spotting prospective business possibilities, assessing the level of services in various locations, and determining economic development policy choices.

## Console Output:

[←](#) Job details [CLONE](#) [DELETE](#) [STOP](#) [REFRESH](#)

Job ID	job-9758633b
Job UUID	0734c546-3494-4166-9de7-e1945a7e49db
Type	Dataproc Job
Status	<span>✓</span> Succeeded

Output [LINE WRAP: OFF](#)

ⓘ Spark jobs take ~60 seconds to initialize resources. [DISMISS](#)

```
23/03/20 22:14:13 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
Top 10 cities with most businesses
+-----+-----+
|city      |count|
+-----+-----+
|Philadelphia|10694|
|Tampa     |7049 |
|Tucson    |6955 |
|Indianapolis|5775 |
|Nashville |5337 |
|New Orleans|4591 |
|Reno      |4209 |
|Saint Louis|3622 |
|Santa Barbara|2639 |
|Boise     |2226 |
+-----+-----+

City with the highest average rating Santa Barbara -> 4.047745358090186
City with the lowest average rating Indianapolis -> 3.607878787878788
23/03/20 22:14:36 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@790770b2{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

EQUIVALENT COMMAND LINE

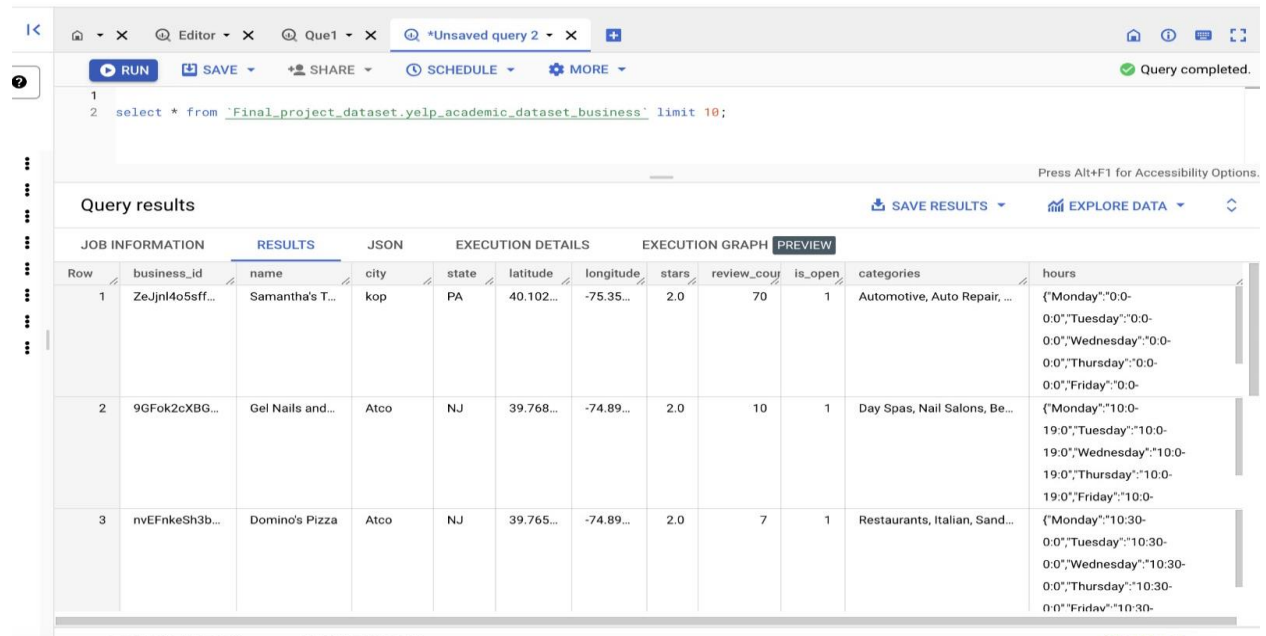
## **DESCRIPTION OF THE PROCESS: -**

Before uploading data to a big query, the data needs to be cleaned first and for that Google cloud Tool-data prep is required to erase those errors. Firstly, on data prep, I select the google cloud storage bucket which contains the business and review files, and imported them to data prep. After that, I created a data flow for this. For cleaning, I used different recipes to clean those noisy data. Then I created a big query table where I uploaded this clean data so that I can perform queries on those data along with that I also delete the CSV file as it was not required and followed the steps which were shown during the assignment.

For performing the analysis through Pyspark, Dataproc is required. Firstly, I uploaded a .py file to the bucket. In the starter code, which was provided during the assignment, I changed the location of the files as well as the bucket and the project name also had a code for connecting Big Query to pyspark. Then I created a cluster by doing some necessary configuration as learned from the assignments. After that, for doing the analysis and generating a result we need to run a job. Now to run a job, we need to give the necessary parameters such as the name of the job, the URI of the .py file that we uploaded in the bucket, and the jar file which was given in the assignment. After this, I was able to successfully run the job and generate the desired result.

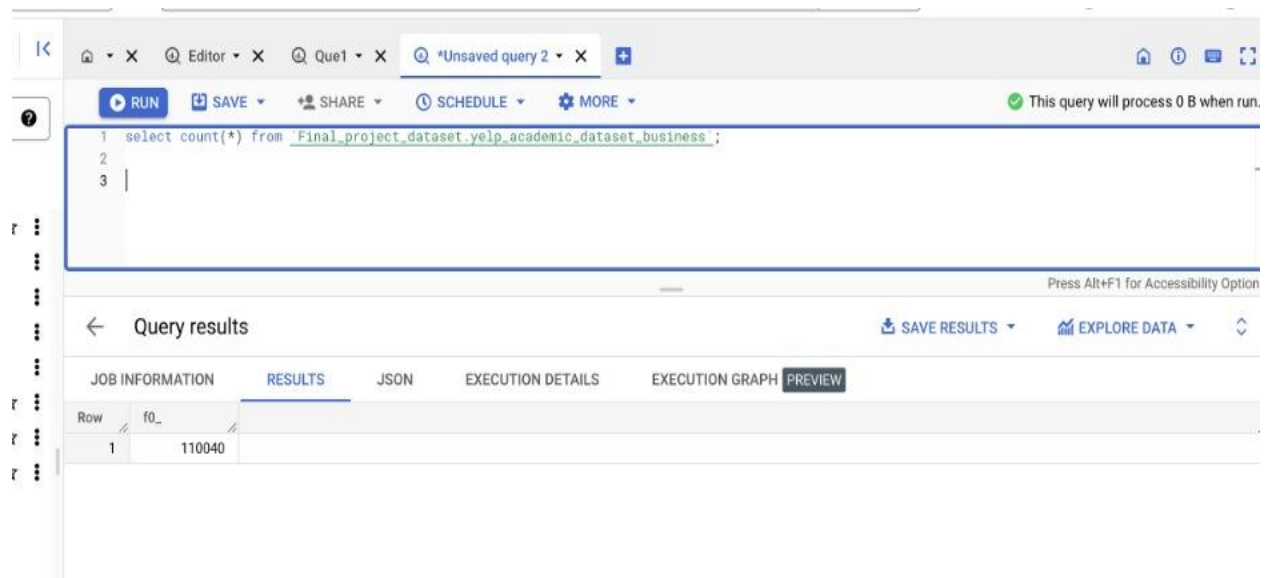
## BIG QUERY PREVIEW SCREENSHOTS:

### Business:



The screenshot shows a query editor interface with a toolbar at the top containing buttons for RUN, SAVE, SHARE, SCHEDULE, and MORE. The query editor displays a SQL query: `select * from 'Final_project_dataset_yelp_academic_dataset_business' limit 10;`. Below the query editor, the results are displayed in a table format. The table has columns for business\_id, name, city, state, latitude, longitude, stars, review\_count, is\_open, categories, and hours. The results show three rows of data for businesses in Atco, NJ.

Row	business_id	name	city	state	latitude	longitude	stars	review_count	is_open	categories	hours
1	ZeJjn14o5eff...	Samantha's T...	kop	PA	40.102...	-75.35...	2.0	70	1	Automotive, Auto Repair, ...	{ "Monday": "0:0-0:0", "Tuesday": "0:0-0:0", "Wednesday": "0:0-0:0", "Thursday": "0:0-0:0", "Friday": "0:0-0:0" }
2	9GFok2cXBG...	Gel Nails and...	Atco	NJ	39.768...	-74.89...	2.0	10	1	Day Spas, Nail Salons, Be...	{ "Monday": "10:0-19:0", "Tuesday": "10:0-19:0", "Wednesday": "10:0-19:0", "Thursday": "10:0-19:0", "Friday": "10:0-19:0" }
3	nvEFnkeSh3b...	Dominos's Pizza	Atco	NJ	39.765...	-74.89...	2.0	7	1	Restaurants, Italian, Sand...	{ "Monday": "10:30-0:0", "Tuesday": "10:30-0:0", "Wednesday": "10:30-0:0", "Thursday": "10:30-0:0", "Friday": "10:30-0:0" }



The screenshot shows a query editor interface with a toolbar at the top containing buttons for RUN, SAVE, SHARE, SCHEDULE, and MORE. The query editor displays a SQL query: `select count(*) from 'Final_project_dataset_yelp_academic_dataset_business';`. Below the query editor, the results are displayed in a table format. The table has columns for Row and f0\_ (representing the count). The results show one row with a count of 110040.

Row	f0_
1	110040

## Review:

The screenshot shows a query editor interface with a tab labeled '\*Unsaved query 2'. The query is:

```
1  
2 select count(*) from 'Final_project_dataset.1_review';
```

Below the query, the 'Query results' section is displayed. It has tabs for 'JOB INFORMATION', 'RESULTS', 'JSON', 'EXECUTION DETAILS', and 'EXECUTION GRAPH'. The 'RESULTS' tab is active, showing a table with two columns: 'Row' and 'f0\_'. The first row contains the value 509665.

Row	f0_
1	509665

The screenshot shows a query editor interface with a tab labeled '\*Unsaved query 2'. The query is:

```
1  
2 select * from 'Final_project_dataset.1_review' limit 10;
```

Below the query, the 'Query results' section is displayed. It has tabs for 'JOB INFORMATION', 'RESULTS', 'JSON', 'EXECUTION DETAILS', and 'EXECUTION GRAPH'. The 'RESULTS' tab is active, showing a table with five columns: 'Row', 'business\_id', 'stars', 'useful', and 'review\_text'. The first four rows are visible.

Row	business_id	stars	useful	review_text
1	-ZVrH2X2QXBfCilbirsw	4	0	The classic Italian hoagie is fantastic and a great value Loved it
2	-ZVrH2X2QXBfCilbirsw	4	0	This place is sadly perm closed I was hoping not however the phone is now disconnected
3	-ZVrH2X2QXBfCilbirsw	4	0	Moving into our new house and I think the Italian hoagie saved my life Happy to be living close
4	-ZVrH2X2QXBfCilbirsw	4	0	Delicious FRESH Good prices Now my one and only hoagie pit stop

I will be adding python script in a zip file which I will be attaching.

## Spark in parallelized computation: -

To use spark for parallelized computation, I created a spark session using:

```
spark = SparkSession \  
    .builder \  
    .master('yarn') \  
    .appName('Yelp_Businesses') \  
    .getOrCreate()
```

Now for loading the data into spark, I used:

```
table_data = sc.newAPIHadoopRDD(  
    'com.google.cloud.hadoop.io.bigquery.JsonTextBigQueryInputFormat',  
    'org.apache.hadoop.io.LongWritable',  
    'com.google.gson.JsonObject',  
    conf = conf)
```

this will reads data from Hadoop Input Format and returns an RDD.

After that, I define a schema for the data frame using the 'StructType' class which defines the column names and data types for the data frame.

```
schema = StructType([  
    # StructField("address", StringType(), True),  
    # StructField("attributes", MapType(StringType(), StringType()), True),  
    StructField("business_id", StringType(), True),  
    StructField("city", StringType(), True),  
    StructField("hours", StringType(), True),  
    StructField("is_open", IntegerType(), True),  
    StructField("latitude", FloatType(), True),  
    StructField("longitude", FloatType(), True),  
    StructField("name", StringType(), True),  
    # StructField("postal_code", StringType(), True),  
    StructField("review_count", IntegerType(), True),  
    StructField("stars", FloatType(), True),  
    StructField("categories", StringType(), True),  
    StructField("state", StringType(), True)  
])
```

### **Statement of Originality: -**

I have collaborated with my group member Abhishek Patel for questions 1 & 2 however; we both performed this analysis on our own individual machines and are able to understand the concept behind it. Question 3 was performed by me.