Abhishek Patel

CS 512

March 20th, 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 business sector is popular in the US and which ones are not, the no. of businesses that is open on Sunday statewise & total no. of businesses which are highly rated category wise.

OSEMN PROCESS: -

Obtain: -

The data for the project is obtained from the yelp dataset which is the same dataset that we worked for our mid-term project. This dataset contains information about the various types of businesses, users, reviews, check-ins, and tips. 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, I divided the initial JSON files to smaller files and edited them using virtual studio code to the standardized JSON format. Then, I uploaded these JSON files to the Google Cloud Storage bucket which can be imported to Dataprep for data wrangling process. In Dataprep, I prepared recipes for handling missing or null values, mismatched values and any outliers in the data that may impact accuracy of results. After the dataset was cleaned, I ran the job and sent it to BigQuery to proceed.

Explore: -

In this stage, we will explore this data. After successfully loading the dataset into BigQuery, the tables business and review were created. I performed insightful analysis on these tables where I answered 3 questions out of which 2 were answered by using PySpark and 1 was answered using BigQuery along with Looker Studio visualization.

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 my dataset as the focus area is to get solutions to 3 data analysis questions using PySpark analysis and visualizations.

iNterpret

This is the final stage of OSEMN. In this stage, I am performing analysis on the dataset using PySpark and Looker Studio visualization. With this analysis, I was able to find the top 10 most reviewed restaurants in Philadelphia and what are the ratings for all these restaurants. It could be understood that the most reviewed restaurants are popular as they were highly rated. I also found out the total no. of businesses statewise that are open on Sunday and are highly rated. This tells about the state and its popular weekend culture.

Overview/Description of the Yelp 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 includes business, review, check-in, tip and user. This dataset as JSON files is not in the standard JSON format (missing the outer major object and missing commas to separate the list of objects) and there is some manual work that needs to be done with the initial data.

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-WlRu9Lw",

        "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

}
```

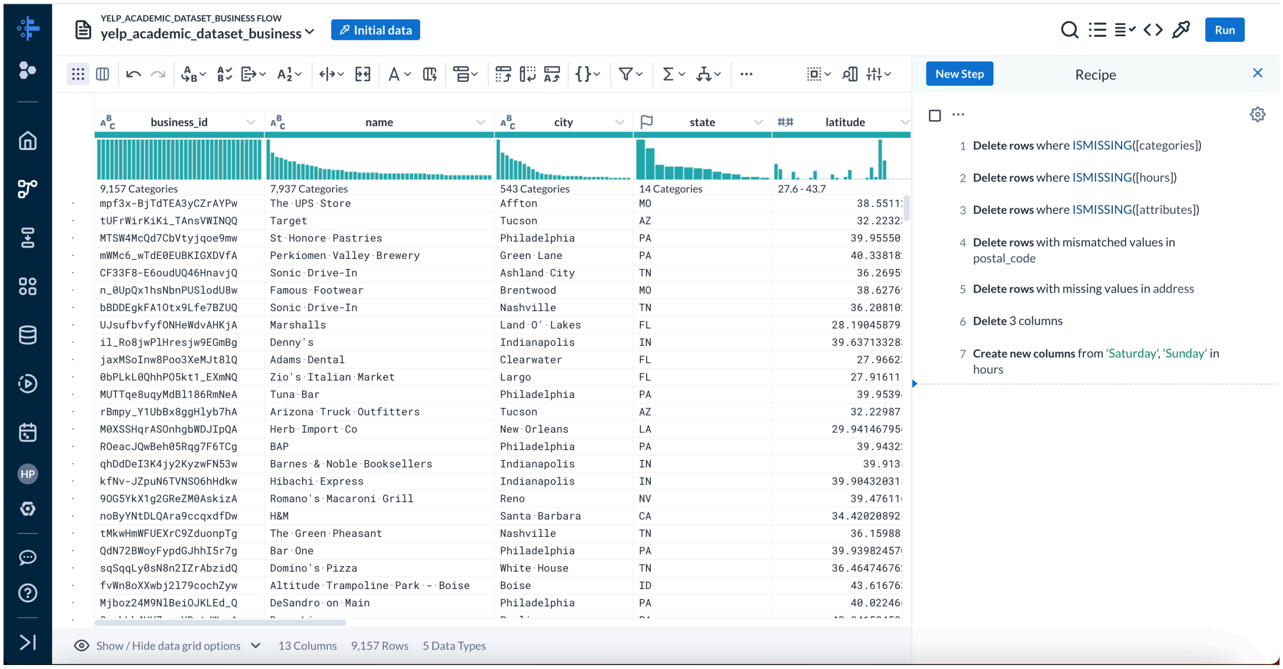
```
{
….
}
```

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, ll63altLtfOgVhEM0KlTqA",
        "fans":1,
        "average_stars":4.15,
        "compliment_hot":0,
        "compliment_note":6,
        "compliment_cool":2,
        "compliment_funny":2,
        "compliment_writer":1,
        "compliment_photos":0
}
{
….
}
```

Overview of Data Wrangling Process:

- First, as you can see the initial data that we started with was not in standard format. So, I edited all the .json files using VSC manually by putting "[,]" for outer main JSON object and "," between JSON objects to convert them to standardized JSON format.
- Secondly, after these files were loaded in Google Cloud Storage, I imported the business.json file to Dataprep where I cleaned the data by using recipe.
- In Dataprep, I took care of the missing and null values in columns such as 'attributes', 'hours', 'categories', and 'address' by removing them.
- I also removed those rows from data that had mismatched values in 'postal_code' column.
- I deleted the columns that were not of any use for my analysis such as 'address', 'postal_code' and 'attributes'. (See below snip)



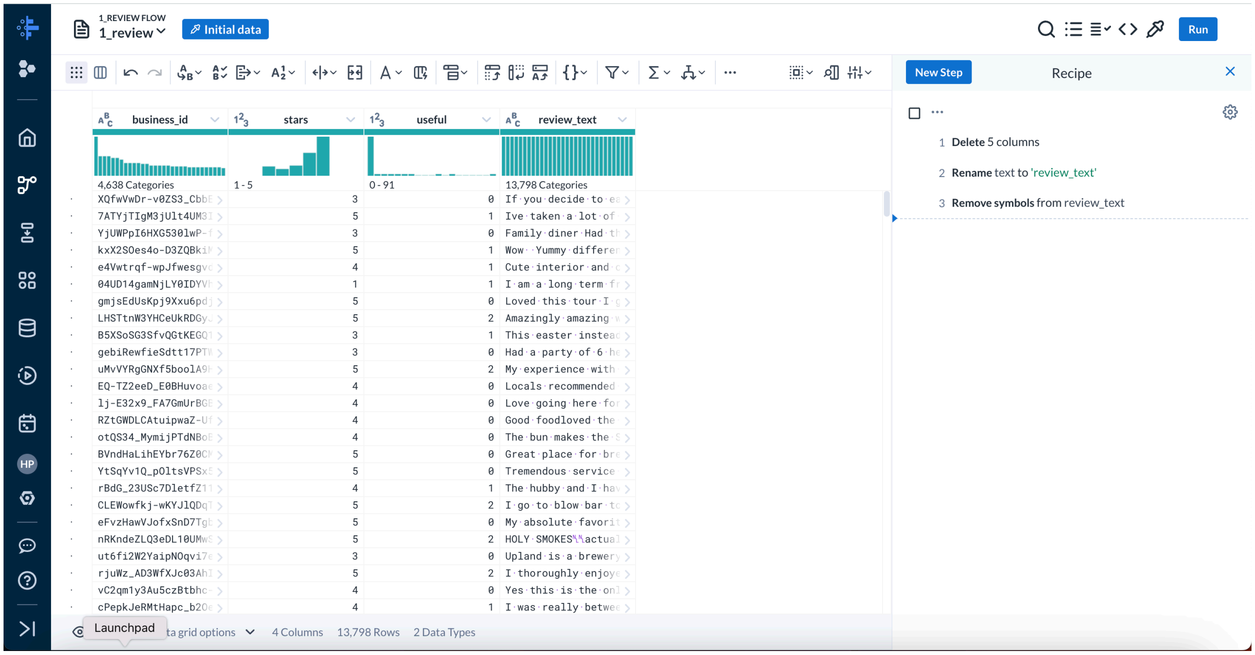
YELP_ACADEMIC_DATASET_BUSINESS FLOW
yelp_academic_dataset_business
Initial data
Run
New Step
Recipe
HP

business_id	name	city	state	latitude
9,157 Categories	7,937 Categories	543 Categories	14 Categories	27.6 - 43.7
mpf3x-BjTdTEA3yCZrAYPw	The UPS Store	Affton	MO	38.5511
tUFrWirKiKi_TAnsVWINQQ	Target	Tucson	AZ	32.2232
MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	Philadelphia	PA	39.95550
mWMc6_wTdE0EUBKIGXDVfA	Perkiomen Valley Brewery	Green Lane	PA	40.33818
CF33F8-E6oudUQ46HnavjQ	Sonic Drive-In	Ashland City	TN	36.2695
n_0UpQx1hsNbnPUSlodU8w	Famous Footwear	Brentwood	MO	38.6276
bBDDEgkFA1Otx9Lfe7BZUQ	Sonic Drive-In	Nashville	TN	36.20810
UJsufbvfyfONHeWdvAHKjA	Marshalls	Land O' Lakes	FL	28.19045879
il_Ro8jwPlHresjw9EGmBg	Denny's	Indianapolis	IN	39.63713328
jaxMSoInw8Poo3XeMJt8lQ	Adams Dental	Clearwater	FL	27.9662
0bPLkL0QhhPO5kt1_EXmNQ	Zio's Italian Market	Largo	FL	27.91611
MUTTqe8uqyMdBl186RmNeA	Tuna Bar	Philadelphia	PA	39.9539
rBmpy_Y1UbBx8ggHlyb7hA	Arizona Truck Outfitters	Tucson	AZ	32.22987
M0XSSHqrASOnhgbWDJIpQA	Herb Import Co	New Orleans	LA	29.94146795
ROeacJQwBeh05Rqg7F6TCg	BAP	Philadelphia	PA	39.9432
qhDdDeI3K4jy2KyzwFN53w	Barnes & Noble Booksellers	Indianapolis	IN	39.913
kfNv-JZpuN6TVNSO6hHdkw	Hibachi Express	Indianapolis	IN	39.90432031
9OG5YkX1g2GReZM0AskizA	Romano's Macaroni Grill	Reno	NV	39.47611
noByYNtDLQAra9ccqxdfDw	H&M	Santa Barbara	CA	34.42020892
tMkwHmWFUEXrC9ZduonpTg	The Green Pheasant	Nashville	TN	36.15988
QdN72BWoyFypdGJhhI5r7g	Bar One	Philadelphia	PA	39.93982457
sqSqqLy0sN8n2IZrAbzidQ	Domino's Pizza	White House	TN	36.46474676
fvWn8oXXwbj2l79cochZyw	Altitude Trampoline Park - Boise	Boise	ID	43.61676
Mjboz24M9NlBeiOJKLEd_Q	DeSandro on Main	Philadelphia	PA	40.02246

Show / Hide data grid options 13 Columns 9,157 Rows 5 Data Types

1 **Delete rows** where ISMISSING([categories])
2 **Delete rows** where ISMISSING([hours])
3 **Delete rows** where ISMISSING([attributes])
4 **Delete rows** with mismatched values in postal_code
5 **Delete rows** with missing values in address
6 **Delete** 3 columns
7 **Create new columns** from 'Saturday', 'Sunday' in hours

- I also imported review.json file and cleaned it by deleting unnecessary columns and I removed symbols from the 'text' column for further analysis. (See below snip)



business_id	stars	useful	review_text
4,638 Categories	1 - 5	0 - 91	13,798 Categories
XQfwVwDr-v0ZS3_CbbE	3	0	If you decide to ea
7ATYjTIgM3jUlt4UM3I	5	1	Ive taken a lot of
YjUWPpI6HXG530lwP-f	3	0	Family diner Had th
kxX2SOes4o-D3ZQBkiM	5	1	Wow Yummy differen
e4Vwtrqf-wpJfwesgvd	4	1	Cute interior and c
04UD14gamNjLY0IDYVh	1	1	I am a long term fr
gmjsEdUsKpj9Xxu6pdj	5	0	Loved this tour I g
LHSTtnW3YHCeUkRDGyJ	5	2	Amazingly amazing w
B5XSoSG3SfvQGtKEGQ1	3	1	This easter instead
gebiRewfieSdtt17PTW	3	0	Had a party of 6 he
uMvVYRgGNXf5boolA9H	5	2	My experience with
EQ-TZ2eeD_E0BHuvoae	4	0	Locals recommended
lj-E32x9_FA7GmUrBGE	4	0	Love going here for
RZtGWDLCAtuipwaZ-Uf	4	0	Good foodloved the
otQS34_MymijPTdNBoE	4	0	The bun makes the S
BVndHaLihEYbr76Z0CM	5	0	Great place for bre
YtSqYv1Q_pOltsVPSx5	5	0	Tremendous service
rBdG_23USc7DletfZ11	4	1	The hubby and I hav
CLEWowfkj-wKYJlQDqT	5	2	I go to blow bar to
eFvzHawVJofxSnD7Tgb	5	0	My absolute favorit
nRKndeZLQ3eDL10UMwS	5	2	HOLY SMOKES actual
ut6fi2W2YaipNOqvi7e	3	0	Upland is a brewery
rjuWz_AD3WfXJc03AhI	5	2	I thoroughly enjoye
vC2qm1y3Au5czBtbhc-	4	0	Yes this is the onl
cPepkJeRMtHapc_b20e	4	1	I was really betwee

ta grid options 4 Columns 13,798 Rows 2 Data Types

1 **Delete** 5 columns

2 **Rename** text to 'review_text'

3 **Remove symbols** from review_text

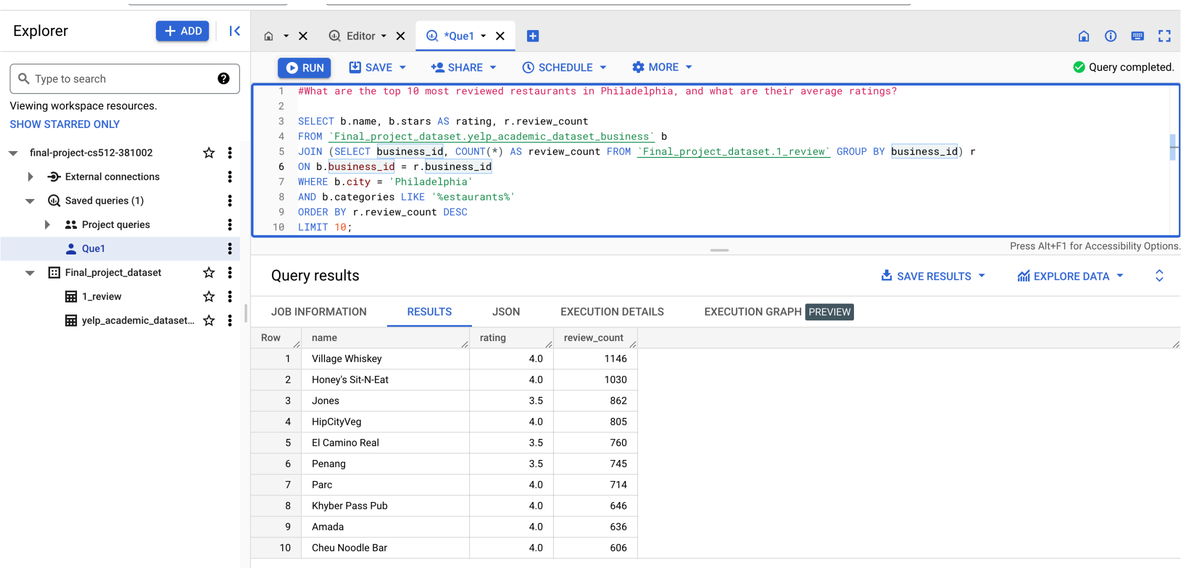
- Finally, my data required for analysis was cleaned and ready to work on.

Data Analysis Questions:

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

   Process to solve:

- To answer this analysis question, I needed business and review tables in BigQuery.
- I am using BigQuery SQL query to analyze the data. I performed JOIN on both these tables.
- I found the results for city = ‘Philadelphia’ and kept the business category = ‘Restaurants’.
- To find the top 10 entries, I performed order by to the total review counts and displayed my result.



```sql
#What are the top 10 most reviewed restaurants in Philadelphia, and what are their average ratings?

SELECT b.name, b.stars AS rating, r.review_count
FROM `Final_project_dataset.yelp_academic_dataset_business` b
JOIN (SELECT business_id, COUNT(*) AS review_count FROM `Final_project_dataset.1_review` GROUP BY business_id) r
ON b.business_id = r.business_id
WHERE b.city = 'Philadelphia'
AND b.categories LIKE '%estaurants%'
ORDER BY r.review_count DESC
LIMIT 10;
```

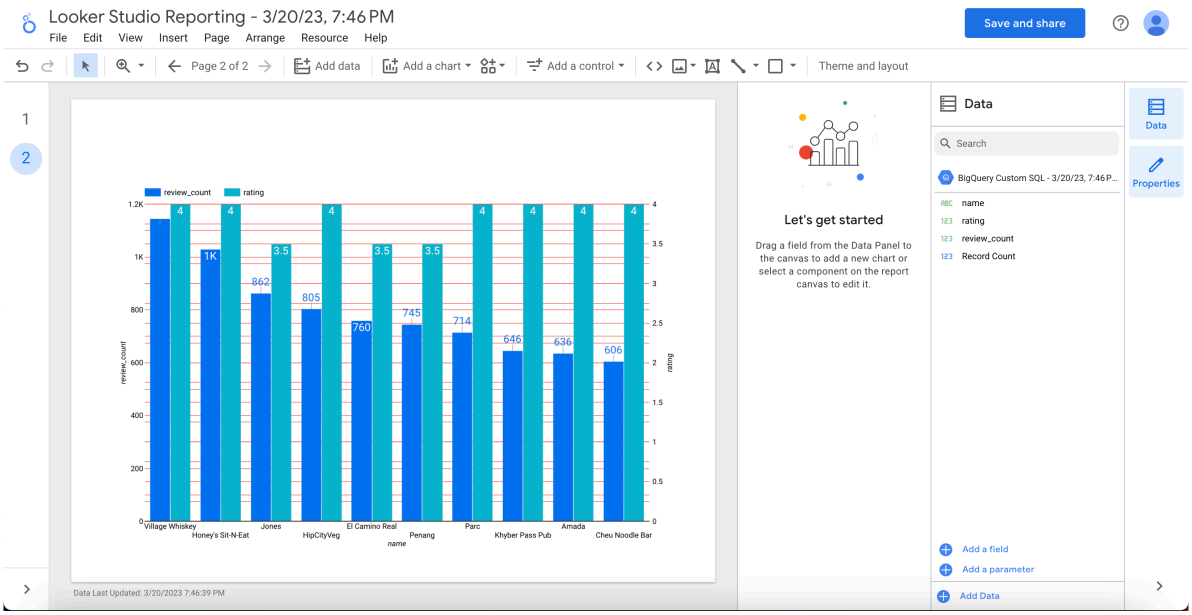
Query results

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

Answer:

This analysis provides information on the most reviewed restaurants in city of Philadelphia which are highly rated. It is observed that the restaurants which are highly rated have many reviews given by customers. This can help the restaurants to evaluate their popularity and competition in any region.

Looker Studio Visualization:



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

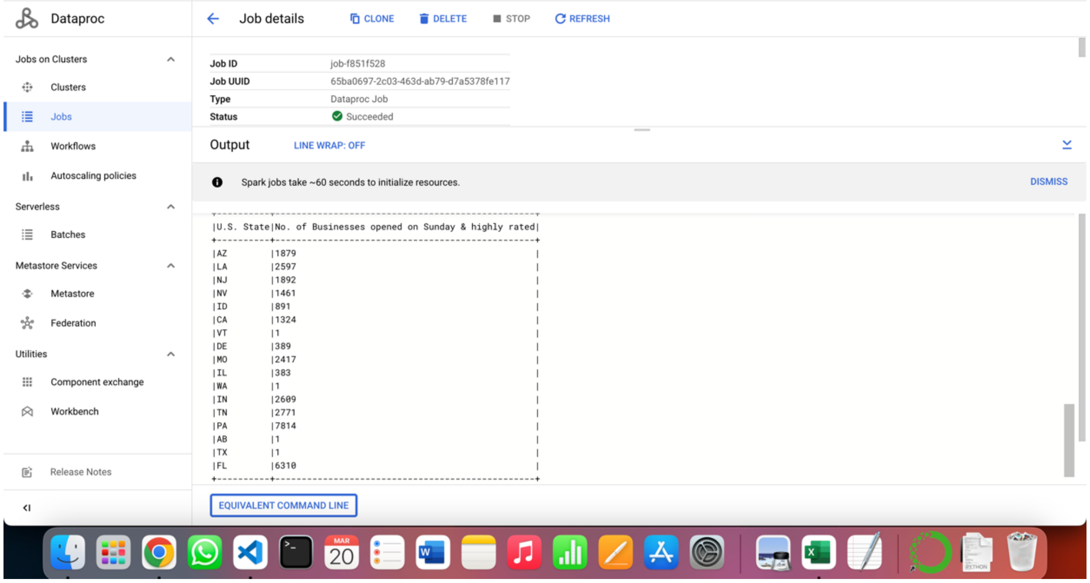
   Process to solve:

   - To answer this question, I took help of the starter code provided by Professor during the Spark planes assignment. It was useful for making connections between PySpark, BigQuery and Google Cloud Storage.
   - To access the value of Sunday in 'hours' column, I converted the data type of 'hours' column to dictionary.
   - I filtered the dataframe that I created from business dataset to only rows having stars>=4 and where the value of Sunday is not null.
   - Then, I grouped by this filtered dataframe by 'state' and found the total counts of businesses statewise.
   - Finally, the results were displayed on the console in DataProc's job screen.

   Answer:

   This analysis provides us with information regarding the total no. of businesses statewise that have their working hours on Sunday and have high ratings. The results obtained here will prove useful for people trying to visit places even on holiday (Sunday) and are high rated.

DataProc Console Output:



Job ID job-f851f528
Job UUID 65ba0697-2c03-463d-ab79-d7a5378fe117
Type Dataproc Job
Status Succeeded

Spark jobs take ~60 seconds to initialize resources.

```
+----------+-----------------------------------------------+
|U.S. State|No. of Businesses opened on Sunday & highly rated|
+----------+-----------------------------------------------+
|AZ        |1879                                           |
|LA        |2597                                           |
|NJ        |1892                                           |
|NV        |1461                                           |
|ID        |891                                            |
|CA        |1324                                           |
|VT        |1                                              |
|DE        |389                                            |
|MO        |2417                                           |
|IL        |383                                            |
|WA        |1                                              |
|IN        |2609                                           |
|TN        |2771                                           |
|PA        |7814                                           |
|AB        |1                                              |
|TX        |1                                              |
|FL        |6310                                           |
+----------+-----------------------------------------------+
```

3. **What is the total no. of businesses that are highly rated category-wise? Find the top 10 categories.**

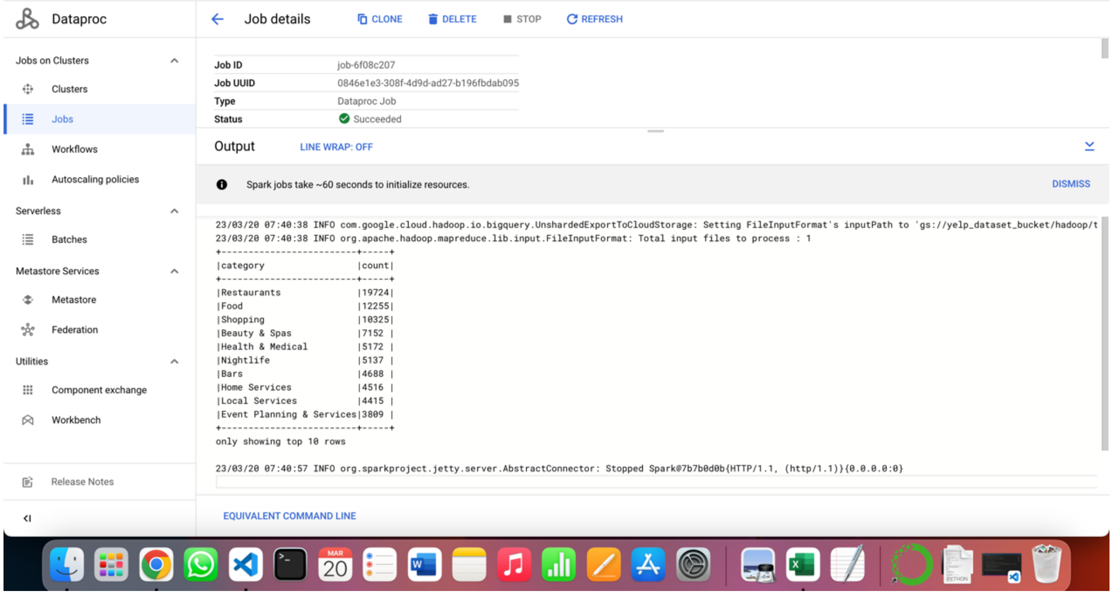
Process to solve:

- To answer this question, I followed the same steps as above till having the starter code and with its help creating a dataframe of business table that is pulled from the BigQuery.
- I filtered out with businesses having high ratings >=4
- Since the question is to find the answer category-wise, I split the categories into each separate category.
- The total no. of businesses are then counted category-wise and the top 10 categories are displayed as result.

Answer:

This analysis provides us with information on top 10 highly rated categories with its total no. of businesses in descending order. This information is useful to understand about which categories are most popular in the US and how these categories are helping towards the country's economy.

DataProc Console Output:



Job ID	job-6f08c207
Job UUID	0846e1e3-308f-4d9d-ad27-b196fbdab095
Type	Dataproc Job
Status	Succeeded

Spark jobs take ~60 seconds to initialize resources.

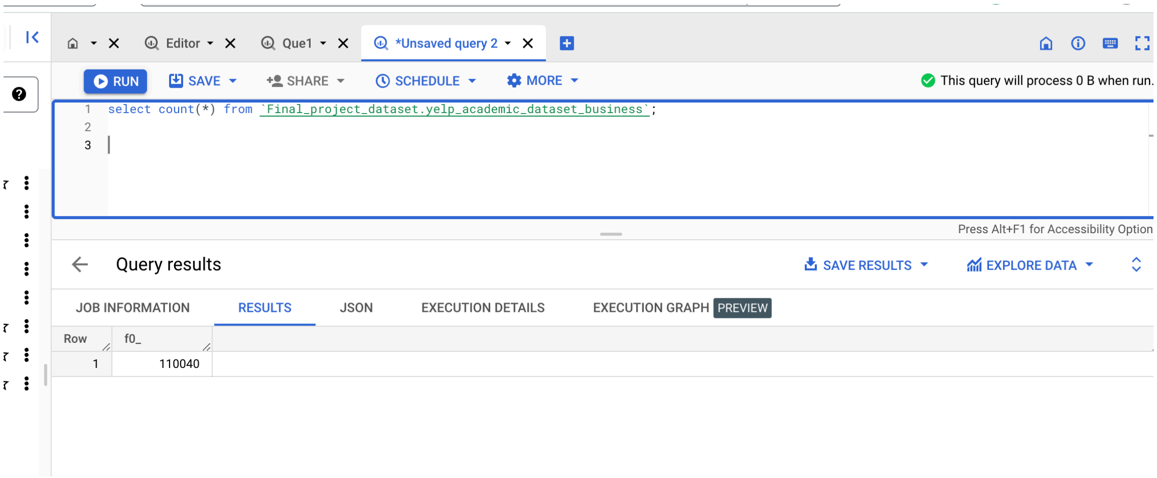
```
23/03/20 07:40:38 INFO com.google.cloud.hadoop.io.bigquery.UnshardedExportToCloudStorage: Setting FileInputFormat's inputPath to 'gs://yelp_dataset_bucket/hadoop/t
23/03/20 07:40:38 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
+------------------------+-----+
|category                |count|
+------------------------+-----+
|Restaurants             |19724|
|Food                    |12255|
|Shopping                |10325|
|Beauty & Spas           |7152 |
|Health & Medical        |5172 |
|Nightlife               |5137 |
|Bars                    |4688 |
|Home Services           |4516 |
|Local Services          |4415 |
|Event Planning & Services|3809 |
+------------------------+-----+
only showing top 10 rows

23/03/20 07:40:57 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@7b7b0d0b{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

Process description to go from original dataset to BigQuery and PySpark.

At first, the data was brought to standard JSON format by adding square brackets and commas between list of JSON objects. Using Google Cloud Tool – Cloud Storage, I uploaded the dataset files which were to be used in DataPrep for data cleaning. In DataPrep, initially I face problems with errors coming as I was importing the large sized files. So, I converted them to small sized files and was successful in importing data in Dataprep. Here, I performed data cleaning using recipes as I explained in the data wrangling process description above. Afterwards, I sent these cleaned data to BigQuery where new tables were created i.e., business and review as I was going to perform analysis on these data.

For performing PySpark analysis, I took help of the starter code given to us earlier in week 8 as we could effectively make connections between PySpark, BigQuery and Google Cloud Storage. I wrote the scripts for question 2 and question 3 and uploaded these .py files to cloud storage. Finally, I created a cluster and run the job there for both these questions that used .py files from cloud storage bucket and gave me the output successfully.

BigQuery Preview Screenshots:

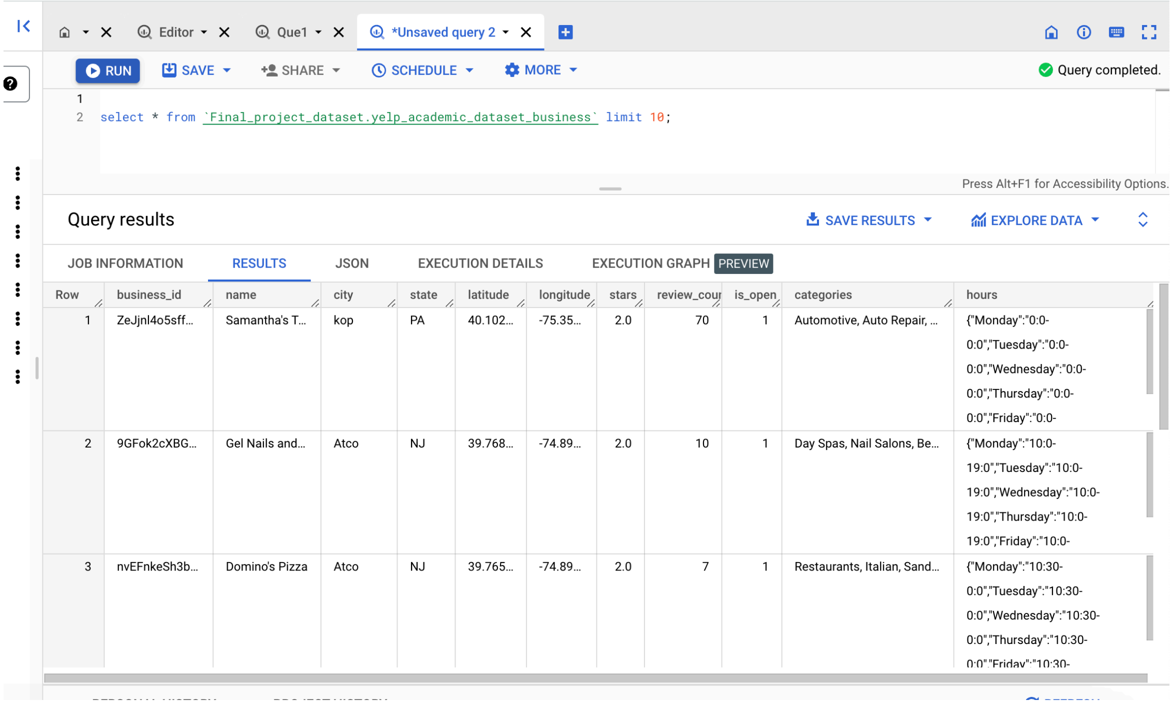


This query will process 0 B when run.

```
select count(*) from `Final_project_dataset.yelp_academic_dataset_business`;

```

Press Alt+F1 for Accessibility Option

Row	f0_
1	110040

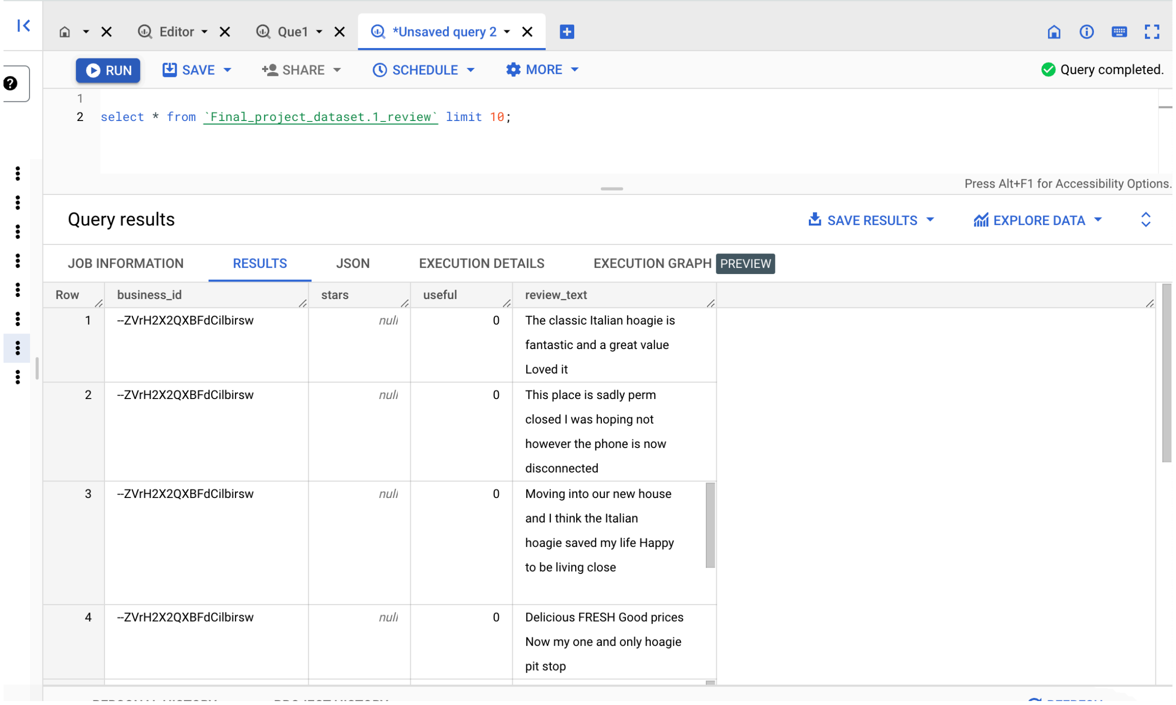


Query completed.

```

select * from `Final_project_dataset.yelp_academic_dataset_business` limit 10;
```

Press Alt+F1 for Accessibility Options.

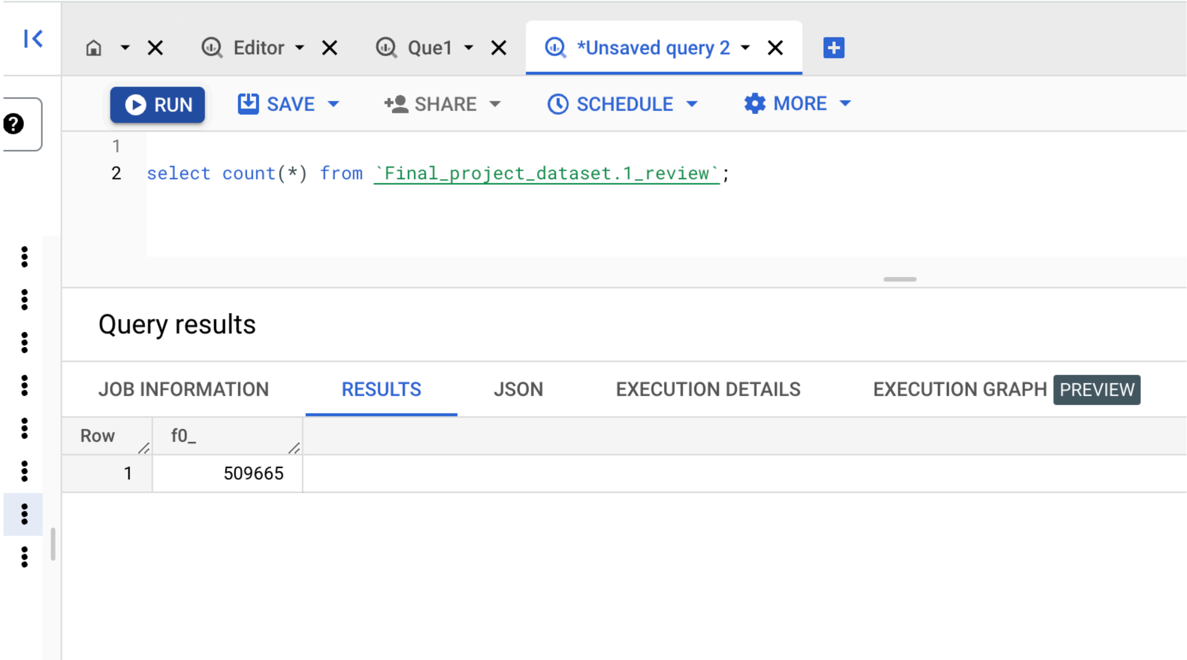
Row	business_id	name	city	state	latitude	longitude	stars	review_cou	is_open	categories	hours
1	ZeJjnl4o5sff...	Samantha's T...	kop	PA	40.102...	-75.35...	2.0	70	1	Automotive, Auto Repair, ...	{"Monday":"0:0-0:0","Tuesday":"0:0-0:0","Wednesday":"0:0-0:0","Thursday":"0:0-0:0","Friday":"0:0-
2	9GFok2cXBG...	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-
3	nvEFnkeSh3b...	Domino'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-



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

Query results

Row	business_id	stars	useful	review_text
1	--ZVrH2X2QXBFdCilbirsw	null	0	The classic Italian hoagie is fantastic and a great value Loved it
2	--ZVrH2X2QXBFdCilbirsw	null	0	This place is sadly perm closed I was hoping not however the phone is now disconnected
3	--ZVrH2X2QXBFdCilbirsw	null	0	Moving into our new house and I think the Italian hoagie saved my life Happy to be living close
4	--ZVrH2X2QXBFdCilbirsw	null	0	Delicious FRESH Good prices Now my one and only hoagie pit stop



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

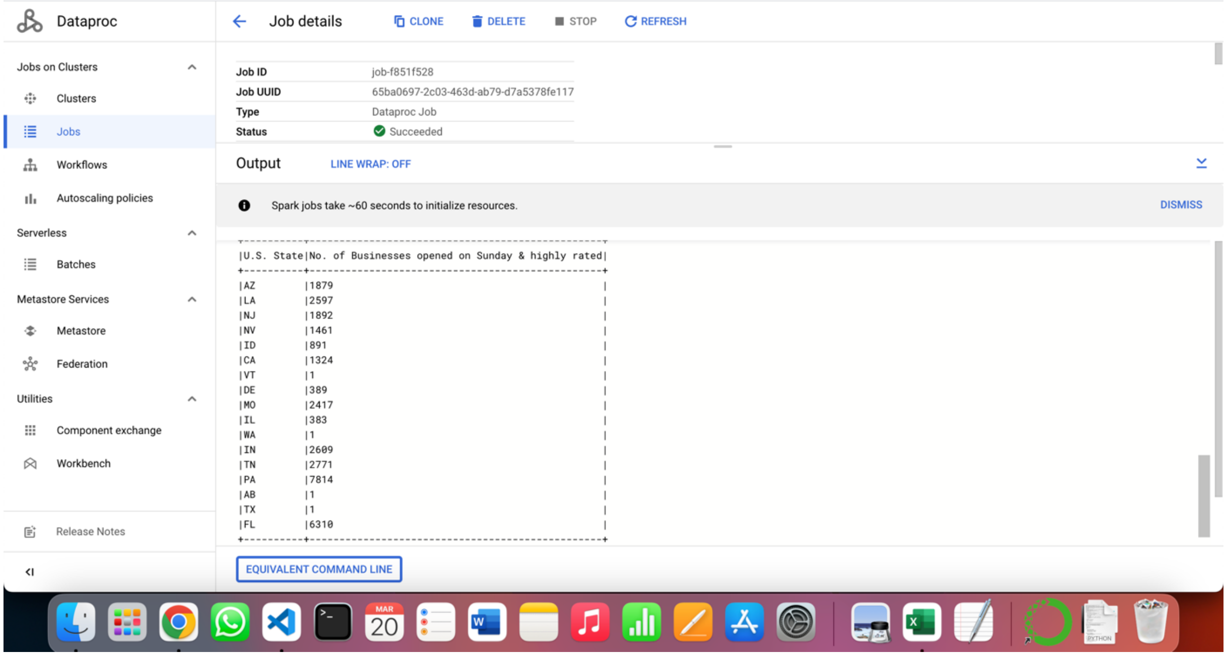
Query results

Row	f0_
1	509665

Python Script for running DataProc Job along with Output:

The Python scripts are attached in the zip folder as instructed.

Below are the screenshots of their output on DataProc:

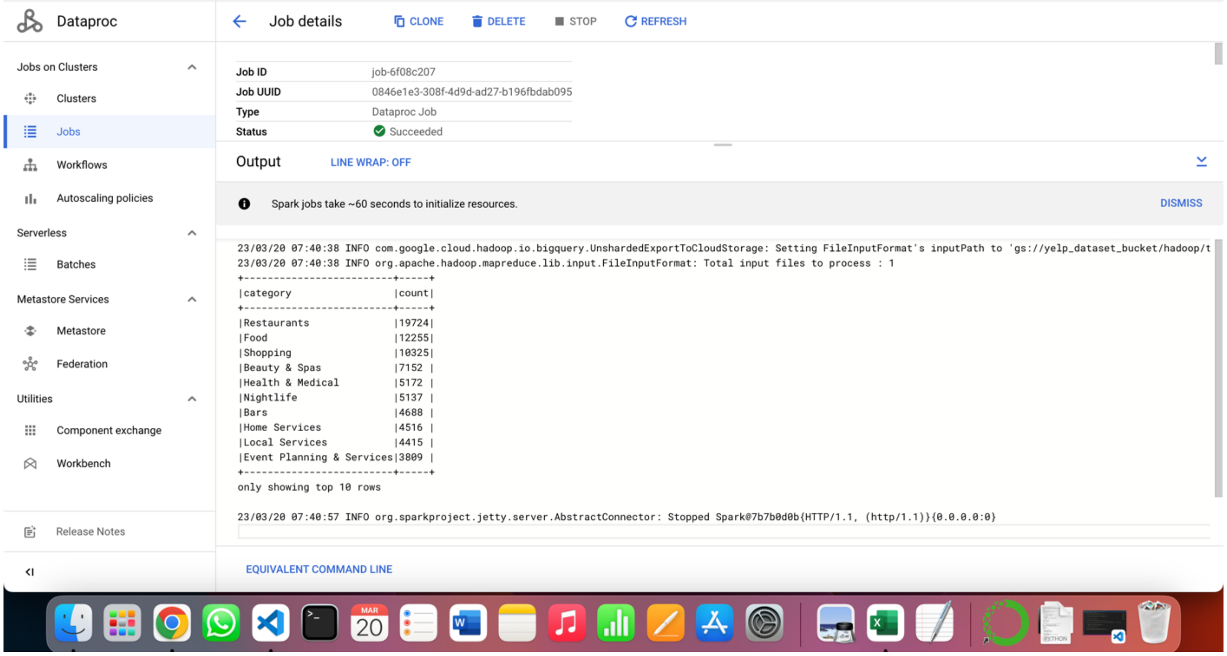


Job ID	job-f851f528
Job UUID	65ba0697-2c03-463d-ab79-d7a5378fe117
Type	Dataproc Job
Status	Succeeded

Output

Spark jobs take ~60 seconds to initialize resources.

```
+----------+----------------------------------------------+
|U.S. State|No. of Businesses opened on Sunday & highly rated|
+----------+----------------------------------------------+
|AZ        |1879                                          |
|LA        |2597                                          |
|NJ        |1892                                          |
|NV        |1461                                          |
|ID        |891                                           |
|CA        |1324                                          |
|VT        |1                                             |
|DE        |389                                           |
|MO        |2417                                          |
|IL        |383                                           |
|WA        |1                                             |
|IN        |2609                                          |
|TN        |2771                                          |
|PA        |7814                                          |
|AB        |1                                             |
|TX        |1                                             |
|FL        |6310                                          |
+----------+----------------------------------------------+
```



Job ID	job-6f08c207
Job UUID	0846e1e3-308f-4d9d-ad27-b196fbdab095
Type	Dataproc Job
Status	Succeeded

Output

Spark jobs take ~60 seconds to initialize resources.

```
23/03/20 07:40:38 INFO com.google.cloud.hadoop.io.bigquery.UnshardedExportToCloudStorage: Setting FileInputFormat's inputPath to 'gs://yelp_dataset_bucket/hadoop/t
23/03/20 07:40:38 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
+-----------------------+-----+
|category               |count|
+-----------------------+-----+
|Restaurants            |19724|
|Food                   |12255|
|Shopping               |10325|
|Beauty & Spas          |7152 |
|Health & Medical       |5172 |
|Nightlife              |5137 |
|Bars                   |4688 |
|Home Services          |4516 |
|Local Services         |4415 |
|Event Planning & Services|3809 |
+-----------------------+-----+
only showing top 10 rows

23/03/20 07:40:57 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@7b7b0d0b{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

Highlight how you used in Parallelized Computation:

I created a spark session to use PySpark for parallelized computation:

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

Now for loading the data into spark, I used:

```
## pull table from big query
table_data = sc.newAPIHadoopRDD(

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

this will read data from Hadoop Input 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
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 collaborated with my group member MohammedSaif for this final project for question 1 and question 2, but all work here is my own. Me and MohammedSaif together solved question1 and question2 by going through previous modules videos. We both performed the solutions on our individual machines after we found the solution. And Question 3 is my own and unique.