

Project Spark

1. Question 1: Create dataframe, load JSON and make temporary table

a. *Load required libraries*

- `from pyspark.sql import SparkSession`
- `from pyspark.sql.types import *`

b. *Create a Spark Session*

- `spark = SparkSession.builder.config("spark.some.config.option", "some-value").getOrCreate()`

c. *Load the JSON file into DataFrame*

- `df=spark.read.json('/user/f201605673498/project_spark/business.json')`

d. *Create a temporary table for the dataframe*

- `df.createOrReplaceTempView("table1")`

2. Question 2: Total business registered with Yelp

a. *To find the total number of business registered with yelp*

- `query2 = spark.sql("select count(business_id) as tot_resturants from table1")`
- `query2.show()`

```
query2.show()
```

```
+-----+
|tot_resturants|
+-----+
|           61184|
+-----+
```

3. Question 3: State-wise count of business registered

a. To find the state-wise count of businesses

- `query3 = spark.sql("select state,count(state) as cnt from table1 group by state order by cnt desc")`
- `query3.count()`

```
In [50]: query3.count()
```

```
Out[50]: 26
```

- `query3.show(20,False)`

```
In [51]: query3.show(20,False)
```

state	cnt
AZ	25230
NV	16485
NC	4963
QC	3921
PA	3041
EDH	2971
WI	2307
BW	934
IL	627
ON	351
SC	189
MLN	123
RP	13
ELN	10
FIF	4
SCB	3
CA	3
XGL	1
NTH	1
HAM	1

only showing top 20 rows

Note: to display all the row `query3.show(query3.count(),False)` is used.

4. Question 4: Display columns which have 5 stars

a. To display name,stars,review_count,city and state column for stars = 5.

- `query4 = spark.sql("select name,stars,review_count,city,state from table1 where stars=5 order by review_count desc")`
- `query4.count()`

```
In [49]: query4.count()
```

```
Out[49]: 7354
```

- `query4.show(20,False)`

```
In [52]: query4.show(20,False)
```

name	stars	review_count	city	state
Art of Flavors	5.0	321	Las Vegas	NV
PNC Park	5.0	306	Pittsburgh	PA
Gaucha Parrilla Argentina	5.0	286	Pittsburgh	PA
Free Vegas Club Passes	5.0	285	Las Vegas	NV
Little Miss BBQ	5.0	267	Phoenix	AZ
Fabulous Eyebrow Threading	5.0	244	Las Vegas	NV
Poke Express	5.0	206	North Las Vegas	NV
Raiding The Rock Vault	5.0	199	Las Vegas	NV
Eco-Tint	5.0	193	Las Vegas	NV
Santos Lucha Libre	5.0	189	Phoenix	AZ
The Whining Pig	5.0	182	Phoenix	AZ
Salon D' Shayn	5.0	165	Scottsdale	AZ
Just-In Time Moving and Delivery	5.0	162	Phoenix	AZ
I Clean Carpets	5.0	147	Phoenix	AZ
Carpet Monkeys Of Las Vegas	5.0	142	Las Vegas	NV
Snow Ono Shave Ice	5.0	139	Las Vegas	NV
A 1 Minute Key Service	5.0	139	Phoenix	AZ
Snapdragon Salon	5.0	138	Phoenix	AZ
Professional Brake Service	5.0	136	Las Vegas	NV
Gelato Dolce Vita	5.0	125	Mesa	AZ

only showing top 20 rows

Note: to display all the row `query4.show(query4.count(),False)` is used.

5. Question 5: Display name, stars, review_count in Las Vegas

a. To display name, stars and review_count column in Las Vegas order by stars and review count.

- `query5 = spark.sql("select name,stars,review_count from table1 where city='Las Vegas' order by stars desc,review_count desc ")`
- `query5.count()`

```
In [53]: query5.count()
```

```
Out[53]: 13601
```

- `query5.show(20,False)`

```
In [44]: query5.show(20,False)
```

name	stars	review_count
Art of Flavors	5.0	321
Free Vegas Club Passes	5.0	285
Fabulous Eyebrow Threading	5.0	244
Raiding The Rock Vault	5.0	199
Eco-Tint	5.0	193
Carpet Monkeys Of Las Vegas	5.0	142
Snow Ono Shave Ice	5.0	139
Professional Brake Service	5.0	136
Platinum Entourage	5.0	124
Battlefield Vegas	5.0	111
Windy City Wash and Wax	5.0	107
Mariscos Playa Escondida	5.0	103
Legacy Air	5.0	97
Umbrella Movers	5.0	94
Boyz II Men	5.0	91
Dreikosen Door Service	5.0	89
Amelia C & Co	5.0	88
CARS Complete Auto Repair Specialists	5.0	87
Josephine Skaught Hairdressing	5.0	86
The Parlor	5.0	84

only showing top 20 rows

Note: to display all the row `query5.show(query5.count(),False)` is used.