%pyspark

from pyspark.sql import SQLContext

sqlCtx = SQLContext(sc)

%pyspark

data = 's3://champagnesbucket/Yelp\_Trends/business/yelp\_academic\_dataset\_business.json'

data2 = 's3://champagnesbucket/Yelp\_Trends/reviews/yelp\_academic\_dataset\_review.json'

data3 = 's3://champagnesbucket/Yelp\_Trends/final/las\_vegas\_business.csv'

data4 = 's3://champagnesbucket/Yelp\_Trends/newdata/checkin\_fn.csv'

data5 = 's3://champagnesbucket/Yelp\_Trends/finalz/business\_data.csv'

%pyspark

# Read Data In

reviews = sqlCtx.read.json(data2)

bus = sqlContext.read.option('inferSchema', 'True').option('Header', 'True').format('csv').load(data3)

pho = sqlContext.read.option('inferSchema', 'True').option('Header', 'True').format('csv').load(data4)

df = sqlCtx.read.json(data)

%pyspark

vegas = sqlContext.read.option('inferSchema', 'True').option('Header', 'True').format('csv').load(data5)

%pyspark

# Data Munging

dayreview = pho.select('business\_id','Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat', 'Sun')

df = sqlCtx.read.json(data)

df = df.drop('hours')

df = df.drop('is\_open')

%pyspark

# Temp Tables

df.registerTempTable('yelp')

reviews.registerTempTable('review')

dayreview.registerTempTable('day')

vegas.registerTempTable('vegas')

%sql

select city, count(city) as Count\_of\_Resturants from yelp group by city order by count(city) DESC limit 10

%sql

select categories, avg(stars) as Avg\_Stars, count(stars) as Count from vegas group by categories order by count(stars) DESC

%sql

select categories, avg(stars) as Avg\_Stars, count(stars) as Count from vegas group by categories order by count(stars) DESC Limit 8

%sql

select postal\_code, avg(stars) as Avg\_Stars from vegas where categories = 'chinese' group by postal\_code order by count(stars)

%sql

select postal\_code, avg(stars) as Avg\_Stars from vegas where categories = ${'american'} group by postal\_code order by count(stars)

%sql

select name, count(text) as Count, avg(R.stars) Avg\_Stars from vegas join review as R using (business\_id) where (postal\_code = '89119' or postal\_code = '89109') and categories = 'chinese' group by name